KINECT-BASED MULTIMODAL GESTURE RECOGNITION USING A TWO-PASS FUSION SCHEME

Georgios Pavlakos, Stavros Theodorakis, Vassilis Pitsikalis, Athanasios Katsamanis and Petros Maragos

School of Electrical and Computer Engineering, National Technical University of Athens, Greece.

ABSTRACT

We present a new framework for multimodal gesture recognition that is based on a two-pass fusion scheme. In this, we deal with a demanding Kinect-based multimodal dataset, which was introduced in a recent gesture recognition challenge. We employ multiple modalities, i.e., visual cues, such as colour and depth images, as well as audio, and we specifically extract feature descriptors of the hands' movement, handshape, and audio spectral properties. Based on these features, we statistically train separate unimodal gesture-word models, namely hidden Markov models, explicitly accounting for the dynamics of each modality. Multimodal recognition of unknown gesture sequences is achieved by combining these models in a late, two-pass fusion scheme that exploits a set of unimodally generated n-best recognition hypotheses. The proposed scheme achieves 88.2% gesture recognition accuracy in the Kinect-based multimodal dataset, outperforming all recently published approaches on the same challenging multimodal gesture recognition task.

Index Terms— multimodal gesture recognition, HMMs, speech recognition, multimodal fusion

1. INTRODUCTION

Gestural interfaces have been gaining increasing attention recently [1, 2]. This can be mainly attributed both to recent technological advances, such as the wide spread of depth sensors, and to groundbreaking research since the famous "put that there" [3]. The natural feeling of gesture interaction can be significantly enhanced by the availability of multiple modalities. Static and dynamic gestures, the form of the hand, as well as speech, all together compose an appealing set of modalities for human-computer interaction that offers significant advantages [4]. All the above, pose numerous challenging research issues for the detection of meaningful information in the visual and audio signals, the employment of appropriate features, the building of effective classifiers, and the multimodal combination of multiple information sources [1].

In this context, our goal is the effective detection and recognition of multimodally expressed gestures as performed freely by multiple users. The demanding dataset [5] that inspired this research effort has been recently acquired for the purpose of the multimodal gesture recognition challenge [6]. This comprises multimodal cultural-anthropological gestures of everyday life, in multi-user spontaneous realizations of both spoken and hand-gesture articulations, intermixed with other random and irrelevant hand, body movements and spoken phrases.

In this paper, we present a multimodal recognition system that exploits the colour, depth and audio signals captured by the Kinect sensor. It extracts features for the handshape configuration, the movement of the hands and the speech signal. We then train hidden Markov models (HMM) for each unimodal cue. These statistical models are at a late-stage integrated in a two-pass fusion scheme. This includes a first-pass that is driven by the most reliable modality. In this step we take advantage of a known approach from the speech recognition community: the multiple n-best sentence hypotheses rescoring scheme [7]. We adapt this concept for our case of complementary modalities, and combine it with a second-pass, in which fusion is performed via parallel HMMs [8] on all modalities *given* the best hypothesis of the first-pass. We have found the novel overall framework to outperform the approaches that participated in the recent demanding multimodal challenge [6], as published in the proceedings of the 2013 ACM ICMI workshop, by reaching an accuracy of 88.2% and leading to a relative error reduction of 7.48% over the first ranked team.

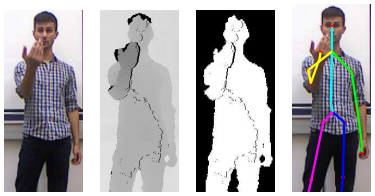
This research work was supported by the European Union under the project "MOBOT" with grant FP7-ICT-2011-9 2.1 - 600796.

2. RELATED WORK

Visual Features: It would be no exaggeration to say that gesture recognition has blossomed since the introduction of depth-based sensors, such as Kinect. Many works face hand tracking by taking advantage of its depth-based tracking (e.g. see works in [5]). Visual cues consist of the movement, position and the shape of the hands. Commonly used features are the 2D/3D center-of-gravity of the hand blob [9], motion features [10], as well as features related with the hand's shape, such as shape moments [9] and Fourier descriptors [11]. Principal component analysis (PCA) is applied for a descriptive representation of handshape, e.g. [12]. Variants of active shape and appearance models are employed for handshape feature extraction [13, 14]. Other approaches employ Histogram of Oriented Gradients (HOG) [15], or scale invariant feature transform (SIFT) [16]. In this work we employ the 3D points of the articulators as extracted from the depth-based skeleton tracking and the HOG descriptors for the handshape cue.

Modeling and HMMs: As far as statistical modeling is concerned, HMMs are employed for the modeling of the dynamics and are applied successfully in hand gesture recognition [17]. Other HMM applications are for instance the threshold model [18] for gesture spotting, and the parametric HMMs [19] for gestures with systematic spatial variation. At the same time Parallel HMMs (PaHMMs) [8] accommodate multiple cues simultaneously, and provide an effective fusion scheme. In this paper we build word-level HMMs both for audio and visual modalities.

N-Best Rescoring and Late Fusion: N-best sentence hypotheses scoring was introduced for the integration of speech and natural language [20], whereas later on it was employed for the integration of different recognition techniques [7]. At the same time, fusion approaches can be broadly classified into early (feature) and late (decision) fusion cases. For the first case, features of the different modal-



(a) RGB (b) Depth (c) Mask (d) Skeleton

Fig. 1. Sample cues. Multi-modal Gesture Challenge 2013 Dataset.

(a) (b) (c) (d)

Fig. 2. (a,b) Arm position variation (low, high) for gesture 'vieni qui'; (c,d) Left and right handed instances of 'vattene'.

ities are early integrated, e.g. by concatenation, and then employed all together for the training of a single multimodal classifier. In the second case separate classifiers are built for each modality and their decisions are late fused. This is usually implemented by combining the class-conditional observation log-likelihoods of each modality into a joint score. Parallel HMMs [8] belong to the second case. In this work we employ the concept of n-best rescoring together with a late fusion parallel HMM framework.

Approaches Evaluated in the Same Dataset: Among the recently published approaches that have been ranked in the first places of the gesture challenge several of them took advantage of the audio modality, whereas for the learning and recognition they employed HMM/GMMs, boosting, random forests, neural networks and support vector machines among others. For a summary see [5]. Wu et al. [21], the first ranked team, are driven by the audio modality based on end-point detection, and then combine classifiers by calculating normalized confidence scores. Authors in [22] are similarly driven by the audio based on a hand-tuned detection algorithm, then they estimate class probabilities per gesture segment and compute their weighted average. Others [23] discard segments not detected in both modalities while employing a temporal overlap coefficient with threshold to merge overlapping modalities' segments. Finally, they recognize the gesture with the highest combined score.

3. MULTIMODAL GESTURE DATASET

Data: The ChaLearn multi-modal gesture challenge dataset [5] provides via Kinect RGB and depth images of face and body, user masks, skeleton information, as well as concurrently recorded audio including the speech utterance accompanying/describing the gesture (see Fig. 1). The vocabulary contains 20 Italian cultural-anthropological gesture-words. The dataset contains three separate sets, namely for development, validation and final evaluation, including 40 users and 13858 gesture-word instances in total.

A challenging task: There is no single way to perform the included cultural gestures, e.g., 'vieni qui' is performed with repeated movements of the hand towards the user, with a variable number of repetitions (see Fig. 2). Similarly, single handed gestures may be performed with either the left or right hand. Further, false alarms are introduced on purpose in both modalities as well as variations in background, lighting, and, resolution, occluded body parts, and different spoken dialects.

4. PROPOSED METHODOLOGY

Our multimodal gesture recognition system essentially implements a two-level approach. First, to independently account for the specificities of each of the modalities involved, separate gesture-word models are trained for speech, skeleton and handshape. These models are then used to generate a set of possible gesture-word sequence hypotheses for a given recording. Then, this original set of hypotheses is multimodally rescored and resorted. Based on the temporal boundaries of the gestures in the best hypothesis, a parallel fusion step exploiting all three modalities further improves recognition.

From a psychobehavioral perspective, gestures and speech are thought to be closely related. They can have complementary or redundant function arising from the same single underlying thought process or mental concept [24]. Gestures convey important communicative information to the listener, but even blind speakers gesture while talking to blind listeners [25].

Gestures in our case occur in parallel with their semantically corresponding speech words. Given the above we assume that the causes of either modality's articulation are the original concepts $C = \{c_i : i = 1, ..., N_C\}$ themselves. The realizations of a concept in each modality m are finally observed in parallel. From our side, we aim to find the underlying common concept given the multimodal observations. Late fusion of the unimodally-based decisions offers a simple and robust way to deconvolve this problem. It allows us before reaching the single best multimodal cause, to get the best unimodal guesses for each modality m, based on the sequence of observations $O_m = [o_{m1}, ...o_{mn}]$, as:

$$\hat{c}_m = \arg\max_{c_i \in C} p(O_m|c_i). \quad (1)$$

Herein we assume there is no prior for the different concepts for either modality.

4.1. Speech, Skeleton and Handshape Modeling

Our modeling methodology essentially follows the keyword-filler paradigm for speech [26, 27] and is based on HMMs. The problem of recognizing a limited number of gesture-words in a video possibly comprising other heterogeneous events as well, is seen as a keyword detection problem. The gesture-words to be recognized are the keywords and all the rest is ignored. Each gesture-word is modeled by an HMM with a common number of states and there is a separate filler HMM to represent all other possible events.

Separate gesture-word models are trained per modality on multiple instances of the gestures as performed by the subjects. The filler model per modality is trained on all training instances. Given these models, recognition hypotheses are generated by means of the Viterbi algorithm [28] on the combined state machine which accepts all possible sequences of gesture-words or filler events.

All our models are left-to-right with Gaussian mixture models (GMMs) representing the state-dependent observation probability distributions. They are initialized by an iterative procedure which sets the model parameters to the mean and covariance of the features in state-corresponding segments of the training instances and refines the segment boundaries via the Viterbi algorithm. Training

is performed using the Baum-Welch algorithm [28], while mixture components are increased incrementally.

4.2. Multimodal Fusion

N-Best Rescoring and Resorting ($P1$): Using the scheme described in the previous section for a single modality and by applying Viterbi decoding we can generate a list of the N-best gesture-word sequence hypotheses $H_1, \ldots, H_N$; N is the number of hypotheses and $H_i = [g_1 g_2 \ldots g_M]$ is a gesture-word sequence. Each hypothesis is accompanied with its corresponding Viterbi score:

$$v_i^m = \max_{q \in Q} \log P(O_m, q | H_i, \lambda), i = 1, \ldots, N, \quad (2)$$

where O_m is the observation sequence for modality m, q is a state sequence of all possible sequences in Q and λ is the corresponding set of models. Given the above hypotheses we rescore them following (2) again but this time employing the HMMs trained for the other modalities along with the corresponding observation sequences. This yields the new hypothesis scores for the rest of the modalities; all scores are then combined into a final score based on which the best hypothesis is chosen.

Currently, the modality providing the initial set of hypotheses is speech since it was found to have the best performance in separate experiments per information stream (Sec. 5). After rescoring the speech hypothesis list based on the handshape and the skeleton we linearly combine their Viterbi scores. The final score of this first pass ($P1$) of fusion for each hypothesis is:

$$v_i^{p1} = \sum_m w_m^{p1} v_i^m \quad (3)$$

where v_i^m is the Viterbi score for hypothesis H_i based on the modality m, and w_m^{p1} is the corresponding weight for the same modality. The stream weights w_m^{p1} are selected in order to optimize the recognition performance in a validation set. The most probable gesture-word sequence after this first fusion step is the one with the maximum combined score.

Second Fusion Pass ($P2$): Herein we exploit the gesture-word level segmentation obtained from the most reliable information stream (in our case, speech). First, we segment the audio, skeleton and handshape observation streams employing the gesture-word level segmentation provided by the best hypothesis generated in the first fusion pass. Segments corresponding to the filler model are ignored. Then for segment s and each modality m we compute the log probability $LL_{s,j}^m = \max_{q \in Q} \log P(O_m, q | \lambda_j^m)$ where λ_j^m are the parameters of the HMM for the gesture-word j and the modality cue m; q is the state sequence. Then we linearly combine the $LL_{s,j}^m$ for all different cues leading to a re-fused log probability of the second fusion pass:

$$LL_{s,j}^{p2} = \sum_m w_m^{p2} LL_{s,j}^m, \quad (4)$$

where w_m^{p2}, is the stream-weight for modality m set to optimize recognition performance in a validation dataset. Finally, the gesture with the maximum score is the recognized one for each segment s.

Single Modalities			Fusion		
Aud.	*Skel.*	*HS*	$P1$	$P2$	$P1 + P2$
78.4	47.6	13.3	85.8	87.2	88.2

Table 1. Single modalities and fusion approaches evaluation. Proposed system's recognition accuracy %, including Audio (Aud.), Skeleton (Skel.), and Handshape (HS).

5. EXPERIMENTS

5.1. Multimodal Features and HMM Parameters

As discussed in Sec. 4.1 we statistically train separate word-gesture level HMMs per modality i.e. audio, skeleton and handshape.

Skeleton Cue: The features employed for the skeleton cue include: the hands' and elbows' 3D position, the hands' 3D position with respect to the corresponding elbow, the 3D direction of the hands' movement, and the 3D distance of hands' centroids. For each gesture we train one left-right HMM using 13 states and 5 mixture components per state.

Handshape Cue: The features employed are HOG as extracted in both hands' segmented images for both RGB and depth modality. We segment the hands by employing the hand's tracking and by performing threshold depth segmentation. Next, for each gesture we train one left-right HMM using 13 states and 1 mixture component per state.

Audio Cue: To efficiently capture the spectral properties of speech signals, our frontend generates 39 acoustic features every 10 msec. Each feature vector comprises 13 Mel Frequency Cepstral Coefficients along with their first and second derivatives. In each left-right HMM we employed 26 states and 6 Gaussians per state. The word insertion penalty was set equal to -400.

In all modalities we built a background HMM (bm) in order to model out-of-vocabulary words. The number of states, mixture components per state, the word insertion penalty in all cases were determined experimentally on the validation set.

5.2. Recognition Results

Single Modalities: In Table 1 we show the recognition results for each modality. As observed the audio modality is the strongest one leading to 78.4% word accuracy in contrast to skeleton and handshape cues which lead to 47.6% and 13.3% respectively.

Separate Fusion Components ($P1$ or $P2$): For the evaluation of the proposed fusion scheme we separately test each component. First, for the $P1$ component we rescore the audio n-best hypothesis list employing all three modalities and linearly combine their scores. Second, the $P2$ component is separately evaluated here, it employs the gesture-word level segmentation of the audio 1-best hypothesis – this is due to the missing first-pass. It then linearly combines the log-likelihood probabilities in each segment.

Two-Pass Fusion $P1 + P2$: We evaluate the proposed scheme by combining sequentially the two components in the two-pass fusion scheme: In detail we first apply the first-pass fusion step ($P1$) leading to the best fused hypothesis as a result of the n-best rescoring. Then follows the $P2$ component as the second-pass fusion step. In this we employ the gesture-word level segmentation of the above best fused hypothesis, leading on the second-pass fused result and the final recognized words.

Results and Comparisons: As shown in Table 1 all the three fusion cases outperform the unimodal cases leading to at least 34.4%

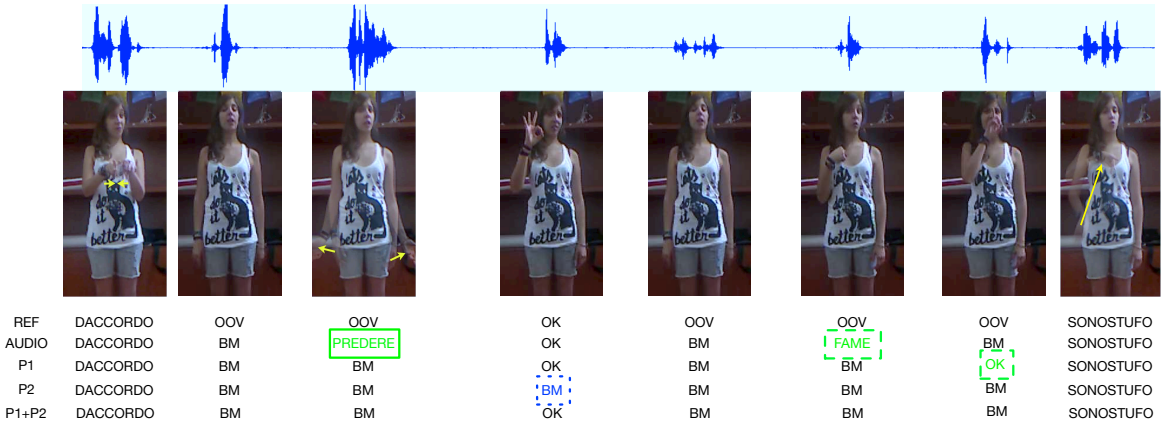


Fig. 3. A decoding word sequence example. Audio (top) and visual modalities (second) via a sequence of images for a word sequence. Ground truth transcriptions ("REF"). Decoding results for the single-audio modality (AUDIO) and the three different fusion schemes (P1, P2 and P1+P2). Errors are highlighted: deletions (blue color) and insertions (green color). A background model (bm) models the out-of-vocabulary (OOV) words.

relative word error rate (WER) reduction[1]. By comparing the separate evaluation of the single fusion components, that is, either $P1$ or $P2$ the $P2$ leads to 9.9% RER compared with $P1$. This is due to the fact that $P1$ is restricted to a single hypothesis out of the unimodal (audio) n-best list. This is in contrast to $P2$ which may recognize a gesture-word sequence that is not present in the audio n-best hypothesis list and fits better to the multi-modal observation vectors. Finally, by comparing the proposed two-pass fusion ($P1+P2$) with $P2$ the former leads to 7% error reduction. This is because in the two-pass fusion scheme the employed gesture-word level segmentation corresponds to the fused hypothesis, that is better than the unimodal (single-audio) 1-best hypothesis in the $P2$ alone. This is clear if we compare the single-audio and $P1$ recognition performances: the latter leads to 34.4% RER.

Example from the Results: A decoding example is shown in Fig. 3. Herein we illustrate both audio and visual modalities for a word sequence accompanied with the ground truth word-level transcriptions (row:"REF"). In addition we show the decoding output employing the single-audio modality (AUDIO) and the three presented fusion cases ($P1$, $P2$ and $P1+P2$). As we observe there are several cases where the subject pronounces an out-of-vocabulary (OOV) word and either perform a gesture or not. This indicates the difficulty of the task as these cases should be ignored. By focusing on the recognized word sequence that employs the single-audio modality we notice two insertions ('PREDERE' and 'FAME'). By employing either the $P1$ or $P2$ the above word insertions are corrected as the visual modality is integrated and helps identifying that these segments correspond to OOV words. Finally, the single pass fusion components lead to errors which the proposed approach manages to deal with: $P1$ causes insertion of "OK", $P2$ of a word deletion "BM". These are in contrast to $P1+P2$ which recognizes correctly the whole sentence.

Comparisons with other approaches in the same task: Herein we compare the recognition results of our proposed multimodal recognition and two-pass fusion framework with other approaches [5] which have been evaluated in the exact recognition task[2]. Among the numerous groups and approaches that participated we list the first four ones as well as the one we submitted during the challenge (pptk). As shown in Table 2 the proposed approach outperforms the others leading to relative error reductions of at least 7.48%. We note that our updated approach from the one submitted during the challenge leads to 31.88% RER. The differential is the following: The fusion scheme employed in this approach was plain $P1$ and moreover the method did not take advantage of all training/validation data during estimation of parameters.

Rank	Approach	Lev. Dist.	Acc.%	RER
-	Our	0.11802	88.198	-
1	iva.mm [21]	0.12756	87.244	+7.48
2	wweight	0.15387	84.613	+23.30
3	E.T. [22]	0.17105	82.895	+31.00
4	MmM	0.17215	82.785	+31.44
5	pptk	0.17325	82.675	+31.88

Table 2. Our approach in comparison with the first 5 places of the Challenge. We include recognition accuracy (Acc.) %, Levenshtein distance (Lev. Dist., see text) and Relative Error Reduction (RER).

6. CONCLUSION

We have presented a framework for kinect based multimodal gesture recognition, exploiting multiple audio and visual modalities. The overall framework is evaluated in a demanding Kinect-based multimodal dataset [5] achieving 88.2% word accuracy. Comparisons include both approaches of several teams that participated in the related challenge, leading to 7.48% relative WER reduction compared to the first ranked team [21], and focused comparisons with other fusion approaches leading to 7% relative WER reduction.

[1] All relative percentages unless stated otherwise refer to relative WER reduction (RER).

[2] In all results presented we follow the same blind testing rules that holded in the challenge, in which we have participated (pptk team). In Table 2 we include for common reference the Levenshtein distance which was also used in the challenge results [5].

7. REFERENCES

[1] A. Jaimes and N. Sebe, "Multimodal human–computer interaction: A survey," *Computer Vision and Image Understanding*, vol. 108, no. 1, pp. 116–134, 2007.

[2] M. Turk, "Multimodal interaction: A review," *Pattern Recognition Letters*, vol. 36, pp. 189–195, 2014.

[3] R. A. Bolt, "Put-that-there: Voice and gesture at the graphics interface," *ACM Computer Graphics*, vol. 14, no. 3, pp. 262–270, 1980.

[4] S. Oviatt and P. Cohen, "Perceptual user interfaces: multimodal interfaces that process what comes naturally," *Communications of the ACM*, vol. 43, no. 3, pp. 45–53, 2000.

[5] S. Escalera, J. Gonzàlez, X. Baró, M. Reyes, O. Lopes, I. Guyon, V. Athistos, and H.J. Escalante, "Multi-modal gesture recognition challenge 2013: Dataset and results," in *Proc. ACM Int'l Conf. on Multimodal Interaction*, 2013, pp. 445–452.

[6] S. Escalera, J. Gonzàlez, X. Baró, M. Reyes, I. Guyon, V. Athitsos, H. Escalante, L. Sigal, A. Argyros, C. Sminchisescu, R. Bowden, and S. Sclaroff, "Chalearn multi-modal gesture recognition 2013: grand challenge and workshop summary," in *Proc. ACM Int'l Conf. on Multimodal Interaction*, 2013, pp. 365–368.

[7] M. Ostendorf, A. Kannan, S. Austin, O. Kimball, R. Schwartz, and J. R. Rohlicek, "Integration of diverse recognition methodologies through reevaluation of n-best sentence hypotheses.," in *Proc. of the Workshop on Speech and Natural Language*, 1991, pp. 83–87.

[8] C. Vogler and D. Metaxas, "A framework for recognizing the simultaneous aspects of American sign language," *Computer Vision and Image Understanding*, vol. 81, pp. 358–384, 2001.

[9] T. Starner, J. Weaver, and A. Pentland, "Real-time American sign language recognition using desk and wearable computer based video," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, no. 12, pp. 1371–1375, 1998.

[10] M. H. Yang, N. Ahuja, and M. Tabb, "Extraction of 2D motion trajectories and its application to hand gesture recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 8, pp. 1061–1074, 2002.

[11] S. Conseil, S. Bourennane, and L. Martin, "Comparison of Fourier descriptors and Hu moments for hand posture recognition," in *Proc. European Conf. on Signal Processing*, 2007, pp. 1960–1964.

[12] W. Du and J. Piater, "Hand modeling and tracking for video-based sign language recognition by robust principal component analysis," in *Proc. ECCV Workshop on Sign, Gesture and Activity*, 2010.

[13] H. Fillbrandt, S. Akyol, and K. F. Kraiss, "Extraction of 3D hand shape and posture from image sequences from sign language recognition," in *Proc. Int'l Conf. on Automatic Face & Gesture Recognition*, 2003, pp. 181–186.

[14] A. Roussos, S. Theodorakis, V. Pitsikalis, and P. Maragos, "Dynamic affine-invariant shape-appearance handshape features and classification in sign language videos," *Journal of Machine Learning Research*, vol. 14, no. 1, pp. 1627–1663, 2013.

[15] P. Buehler, M. Everingham, and A. Zisserman, "Learning sign language by watching TV (using weakly aligned subtitles)," in *Proc. Conf. on Computer Vision & Pattern Recognition*, 2009, pp. 2961–2968.

[16] A. Farhadi, D. Forsyth, and R. White, "Transfer learning in sign language," in *Proc. Conf. on Computer Vision & Pattern Recognition*, 2007, pp. 1–8.

[17] Y. Nam and K. Wohn, "Recognition of space-time hand-gestures using hidden Markov model," in *Proc. ACM Symposium on Virtual Reality Software and Technology*, 1996, pp. 51–58.

[18] H. K. Lee and J. H. Kim, "An HMM-based threshold model approach for gesture recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 21, no. 10, pp. 961–973, 1999.

[19] A. Wilson and A. Bobick, "Parametric hidden Markov models for gesture recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 21, no. 9, pp. 884–900, 1999.

[20] Y. L. Chow and R. Schwartz, "The n-best algorithm: An efficient procedure for finding top n sentence hypotheses," in *Proc. of the Workshop on Speech and Natural Language*, 1989, pp. 199–202.

[21] J. Wu, J. Cheng, C. Zhao, and H. Lu, "Fusing multi-modal features for gesture recognition," in *Proc. ACM Int'l Conf. on Multimodal Interaction*, 2013, pp. 453–460.

[22] I. Bayer and T. Silbermann, "A multi modal approach to gesture recognition from audio and video data," in *Proc. ACM Int'l Conf. on Multimodal Interaction*, 2013, pp. 461–466.

[23] K. Nandakumar, K. W. Wan, S. Chan, W. Ng, J. G. Wang, and W. Y. Yau, "A multi-modal gesture recognition system using audio, video, and skeletal joint data," in *Proc. ACM Int'l Conf. on Multimodal Interaction*, 2013, pp. 475–482.

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

[25] J. M. Iverson and S. Goldin-Meadow, "Why people gesture when they speak," *Nature*, vol. 396, no. 6708, pp. 228–228, 1998.

[26] J. Wilpon, L. R. Rabiner, C. H. Lee, and E. R. Goldman, "Automatic recognition of keywords in unconstrained speech using hidden Markov models," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 38, no. 11, pp. 1870–1878, 1990.

[27] R. C. Rose and D. B. Paul, "A hidden Markov model based keyword recognition system," in *Proc. Int'l Conf. on Acoustics, Speech and Signal Processing*, 1990, pp. 129–132.

[28] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.