Journal of Machine Learning Research 16 (2015) 255-284 Submitted 3/14; Revised 9/14; Published 2/15

Multimodal Gesture Recognition via Multiple Hypotheses Rescoring

Vassilis Pitsikalis VPITSIK@CS.NTUA.GR
Athanasios Katsamanis NKATSAM@CS.NTUA.GR
Stavros Theodorakis STH@CS.NTUA.GR
Petros Maragos MARAGOS@CS.NTUA.GR
National Technical University of Athens
School of Electrical and Computer Engineering
Zografou Campus, Athens 15773, Greece

Editors: Isabelle Guyon, Vassilis Athitsos and Sergio Escalera

Abstract

We present a new framework for multimodal gesture recognition that is based on a multiple hypotheses rescoring fusion scheme. We specifically deal with a demanding Kinect-based multimodal data set, introduced in a recent gesture recognition challenge (ChaLearn 2013), where multiple subjects freely perform multimodal gestures. We employ multiple modalities, that is, visual cues, such as skeleton data, color and depth images, as well as audio, and we extract feature descriptors of the hands' movement, handshape, and audio spectral properties. Using a common hidden Markov model framework we build single-stream gesture models based on which we can generate multiple single stream-based hypotheses for an unknown gesture sequence. By multimodally rescoring these hypotheses via constrained decoding and a weighted combination scheme, we end up with a multimodally-selected best hypothesis. This is further refined by means of parallel fusion of the monomodal gesture models applied at a segmental level. In this setup, accurate gesture modeling is proven to be critical and is facilitated by an activity detection system that is also presented. The overall approach achieves 93.3% gesture recognition accuracy in the ChaLearn Kinect-based multimodal data set, significantly outperforming all recently published approaches on the same challenging multimodal gesture recognition task, providing a relative error rate reduction of at least 47.6%.

Keywords: multimodal gesture recognition, HMMs, speech recognition, multimodal fusion, activity detection

1. Introduction

Human communication and interaction takes advantage of multiple sensory inputs in an impressive way. Despite receiving a significant flow of multimodal signals, especially in the audio and visual modalities, our cross-modal integration ability enables us to effectively perceive the world around us. Examples span a great deal of cases. Cross-modal illusions are indicative of lower perceptual multimodal interaction and plasticity (Shimojo and Shams, 2001): for instance, when watching a video, a sound is perceived as coming from the speakers lips (the ventriloquism effect) while, in addition, speech perception may be affected by whether the lips are visible or not (the McGurk effect).

At a higher level, multimodal integration is also regarded important for language production and this is how the notion of multimodal gestures can be introduced. Several authors, as McNeill (1992), support the position that hand gestures hold a major role, and together with speech they are considered to have a deep relationship and to form an integrated system (Bernardis and Gentilucci, 2006) by interacting at multiple linguistic levels. This integration has been recently explored in terms of communication by means of language comprehension (Kelly et al., 2010). For instance, speakers pronounce words while executing hand gestures that may have redundant or complementary nature, and even blind speakers gesture while talking to blind listeners (Iverson and Goldin-Meadow, 1998). From a developmental point of view, see references in the work of Bernardis and Gentilucci (2006), hand movements occur in parallel during babbling of 6-8 month children, whereas word comprehension at the age of 8-10 months goes together with deictic gestures. All the above suffice to provide indicative evidence from various perspectives that hand gestures and speech seem to be interwoven.

In the area of human-computer interaction gesture has been gaining increasing attention (Turk, 2014). This is attributed both to recent technological advances, such as the wide spread of depth sensors, and to groundbreaking research since the famous "put that there" (Bolt, 1980). 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 that offers significant advantages (Oviatt and Cohen, 2000).

In this context, we focus on the effective detection and recognition of multimodally expressed gestures as performed *freely* by multiple users. Multimodal gesture recognition (MGR) poses numerous challenging research issues, such as detection of meaningful information in audio and visual signals, extraction of appropriate features, building of effective classifiers, and multimodal combination of multiple information sources (Jaimes and Sebe, 2007). The demanding data set (Escalera et al., 2013b) used in our work has been recently acquired for the needs of the multimodal gesture recognition challenge (Escalera et al., 2013a). It comprises multimodal cultural-anthropological gestures of everyday life, in spontaneous realizations of both spoken and hand-gesture articulations by multiple subjects, intermixed with other random and irrelevant hand, body movements and spoken phrases.

A successful multimodal gesture recognition system is expected to exploit both speech and computer vision technologies. Speech technologies and automatic speech recognition (Rabiner and Juang, 1993) have a long history of advancements and can be considered mature when compared to the research challenges found in corresponding computer vision tasks. The latter range from low-level tasks that deal with visual descriptor representations (Li and Allinson, 2008), to more difficult ones, such as recognition of action (Laptev et al., 2008), of facial expressions, handshapes and gestures, and reach higher-level tasks such as sign language recognition (Agris et al., 2008). However, recently the incorporation of depth enabled sensors has assisted to partially overcome the burden of detection and tracking, opening the way for addressing more challenging problems. The study of *multiple modalities' fusion* is one such case, that is linked with subjects discussed above.

Despite the progress seen in either unimodal cases such as the fusion of multiple speech cues for speech recognition (e.g., Bourlard and Dupont, 1997) or the multimodal case of audio-visual speech (Potamianos et al., 2004; Glotin et al., 2001; Papandreou et al., 2009),

the integration of dissimilar cues in MGR poses several challenges; even when several cues are excluded such as facial ones, or the eye gaze. This is due to the complexity of the task that involves several intra-modality diverse cues, as the 3D hands' shape and pose. These require different representations and may occur both sequentially and in parallel, and at different time scales and/or rates. Most of the existing gesture-based systems have certain limitations, for instance, either by only allowing a reduced set of symbolic commands based on simple hand postures or 3D pointing (Jaimes and Sebe, 2007), or by considering single-handed cases in controlled tasks. Such restrictions are indicative of the task's difficulty despite already existing work (Sharma et al., 2003) even before the appearance of depth sensors (Weimer and Ganapathy, 1989).

The fusion of multiple information sources can be either early, late or intermediate, that is, either at the data/feature level, or at the stage of decisions after applying independent unimodal models, or in-between; for further details refer to relative reviews (Jaimes and Sebe, 2007; Maragos et al., 2008). In the case of MGR late fusion is a typical choice since involved modalities may demonstrate synchronization in several ways (Habets et al., 2011) and possibly at higher linguistic levels. This is in contrast, for instance, to the case of combining lip movements with speech in audio-visual speech where early or state-synchronous fusion can be applied, with synchronization at the phoneme-level.

In this paper, we present a multimodal gesture recognition system that exploits the color, depth and audio signals captured by a Kinect sensor. The system first extracts features for the handshape configuration, the movement of the hands and the speech signal. Based on the extracted features and statistically trained models, single modality-based hypotheses are then generated for an unknown gesture sequence. The underlying single-modality modeling scheme is based on gesture-level hidden Markov models (HMMs), as described in Section 3.1. These are accurately initialized by means of a model-based activity detection system for each modality, presented in Section 3.3. The generated hypotheses are re-evaluated using a statistical *multimodal multiple hypotheses fusion* scheme, presented in Section 3.2. The proposed scheme builds on previous work on N-best rescoring: N-best sentence hypotheses scoring was introduced for the integration of speech and natural language by Chow and Schwartz (1989) and has also been employed for the integration of different recognition systems based on the same modality, e.g., by Ostendorf et al. (1991), or for audio-visual speech recognition by Glotin et al. (2001). *Given* the best multimodally-selected hypothesis, and the implied gesture temporal boundaries in all information streams, a final *segmental parallel fusion* step is applied based on parallel HMMs (Vogler and Metaxas, 2001). We show in Section 5 that the proposed overall MGR framework outperforms the approaches that participated in the recent demanding multimodal challenge (Escalera et al., 2013a), as published in the proceedings of the workshop, by reaching an accuracy of 93.3 and leading to a relative error rate (as Levenshtein distance) reduction of 47% over the first-ranked team.

2. Related Work

Despite earlier work in multimodal gesture recognition, it is considered an open field, related to speech recognition, computer vision, gesture recognition and human-computer interaction. As discussed in Section 1 it is a multilevel problem posing challenges on audio

and visual processing, on multimodal stream modeling and fusion. Next, we first consider works related to the recent advances on multimodal recognition, including indicative works evaluated in the same ChaLearn challenge and recognition task by sharing the exact training/testing protocol and data set. Then, we review issues related to basic components and tasks, such as visual detection and tracking, visual representations, temporal segmentation, statistical modeling and fusion.

There are several excellent reviews on multimodal interaction either from the computer vision or human-computer interaction aspect (Jaimes and Sebe, 2007; Turk, 2014). Since earlier pioneering works (Bolt, 1980; Poddar et al., 1998) there has been an explosion of works in the area; this is also due to the introduction of everyday usage depth sensors (e.g., Ren et al., 2011). Such works span a variety of applications such as the recent case of gestures and accompanying speech integration for a problem in geometry (Miki et al., 2014), the integration of nonverbal auditory features with gestures for agreement recognition (Bousmalis et al., 2011), or within the aspect of social signal analysis (Ponce-López et al., 2013); Song et al. (2013) propose a probabilistic extension of first-order logic, integrating multimodal speech/visual data for recognizing complex events such as everyday kitchen activities.

The ChaLearn task is an indicative case of the effort recently placed in the field: Published approaches ranked in the first places of this gesture challenge, employ multimodal signals including audio, color, depth and skeletal information; for learning and recognition one finds approaches ranging from hidden Markov models (HMMs)/Gaussian mixture models (GMMs) to boosting, random forests, neural networks and support vector machines among others. Next, we refer to indicative approaches from therein, (Escalera et al., 2013b). In Section 5 we refer to specific details for the top-ranked approaches that we compare with. Wu et al. (2013), the first-ranked team, are driven by the audio modality based on end-point detection, to detect the multimodal gestures; then they combine classifiers by calculating normalized confidence scores. Bayer and Thierry (2013) are also driven by the audio based on a hand-tuned detection algorithm, then they estimate class probabilities per gesture segment and compute their weighted average. Nandakumar et al. (2013) are driven by both audio HMM segmentation, and skeletal points. They discard segments not detected in both modalities while employing a temporal overlap coefficient to merge overlapping modalities' segments. Finally, they recognize the gesture with the highest combined score. Chen and Koskela (2013) employ the extreme learning machine, a class of single-hidden layer feed-forward neural network and apply both early and late fusion. In a late stage, they use the geometric mean to fuse the classification outputs. Finally, Neverova et al. (2013) propose a multiple-scale learning approach that is applied on both temporal and spatial dimension while employing a recurrent neural network. Our contribution in the specific area of multimodal gestures recognition concerns the employment of a late fusion scheme based on multiple hypothesis rescoring. The proposed system, also employing multimodal activity detectors, all in a HMM statistical framework, demonstrates improved performance over the rest of the approaches that took part in the specific ChaLearn task.

From the visual processing aspect the first issue to be faced is *hand detection* and *tracking*. Regardless of the boost offered after the introduction of depth sensors there are unhandled cases as in the case of low quality video or resolution, in complex scene backgrounds with multiple users, and varying illumination conditions. Features employed are related to skin color, edge information, shape and motion for hand detection (Argyros

and Lourakis, 2004; Yang et al., 2002), and learning algorithms such as boosting (Ong and Bowden, 2004). *Tracking* is based on blobs (Starner et al., 1998; Tanibata et al., 2002; Argyros and Lourakis, 2004), hand appearance (Huang and Jeng, 2001), or hand boundaries (Chen et al., 2003; Cui and Weng, 2000), whereas modeling techniques include Kalman filtering (Binh et al., 2005), the condensation method (Isard and Blake, 1998), or full upper body pose tracking (Shotton et al., 2013). Others directly employ global image features (Bobick and Davis, 2001). Finally, Alon et al. (2009) employ a unified framework that performs spatial segmentation simultaneously with higher level tasks. In this work, similarly to other authors, see works presented by Escalera et al. (2013b), we take advantage of the Kinect-provided skeleton tracking.

Visual feature extraction aims at the representation of the movement, the position and the shape of the hands. Representative measurements include the center-of-gravity of the hand blob (Bauer and Kraiss, 2001), motion features (Yang et al., 2002), as well as features related with the hand's shape, such as shape moments (Starner et al., 1998) or sizes and distances within the hand (Vogler and Metaxas, 2001). The contour of the hand is also used for invariant features, such as Fourier descriptors (Conseil et al., 2007). handshape representations are extracted via principal component analysis (e.g., Du and Piater, 2010), or with variants of active shape and appearance models (Roussos et al., 2013). Other approaches (e.g. Dalal and Triggs, 2005) employ general purpose features as the Histogram of Oriented Gradients (HOG) (Buehler et al., 2009), or the scale invariant feature transform (Lowe, 1999). Li and Allinson (2008) present a review on local features. 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.

Temporal detection or *segmentation* of meaningful information concerns another important aspect of our approach. Often the segmentation problem is seen in terms of gesture spotting, that is, for the detection of the meaningful gestures, as adapted from the case of speech (Wilcox and Bush, 1992) where all non-interesting patterns are modeled by a single filler model. Specifically, Lee and Kim (1999) employ in similar way an ergodic model termed as threshold model to set adaptive likelihood thresholds. Segmentation may be also seen in combination with recognition as by Alon et al. (2009) or Li and Allinson (2007); in the latter, start and end points of gestures are determined by zero crossing of likelihoods' difference between gesture/non-gestures. There has also been substantial related work in sign language tasks: Han et al. (2009) explicitly perform segmentation based on motion discontinuities, Kong and Ranganath (2010) segment trajectories via rule-based segmentation, whereas others apply systematic segmentation as part of the modeling of sub-sign components (sub-units) (Bauer and Kraiss, 2001); the latter can be enhanced by an unsupervised segmentation component (Theodorakis et al., 2014) or by employing linguistic-phonetic information (Pitsikalis et al., 2011), leading to multiple subunit types. In our case, regardless of the availability of ground truth temporal gesture annotations we employ independent monomodal model-based activity detectors that share a common HMM framework. These function independently of the ground truth annotations, and are next exploited at the statistical modeling stage.

Multimodal gesture recognition concerns multiple dynamically varying streams, requiring the handling of multiple variable time-duration diverse cues. Such requirements are met by approaches such as hidden Markov models that have been found to efficiently model

temporal information. The corresponding framework further provides efficient algorithms, such as BaumWelch and Viterbi (Rabiner and Juang, 1993), for evaluation, learning, and decoding. For instance, Nam and Wohn (1996) apply HMMs in gesture recognition, Lee and Kim (1999) in gesture spotting, whereas parametric HMMs (Wilson and Bobick, 1999) are employed for gestures with systematic variation. At the same time parallel HMMs (Vogler and Metaxas, 2001) accommodate multiple cues simultaneously. Extensions include conditional random fields (CRFs) or generalizations (Wang et al., 2006), while non-parametric methods are also present in MGR tasks (Celebi et al., 2013; Hernández-Vela et al., 2013). In this paper we build word-level HMMs, which fit our overall statistical framework, both for audio and visual modalities, while also employing parallel HMMs for late fusion.

3. Proposed Methodology

To better explain the proposed multimodal gesture recognition framework let us first describe a use case. Multimodal gestures are commonly used in various settings and cultures (Morris et al., 1979; Kendon, 2004). Examples include the "OK" gesture expressed by creating a circle using the thumb and forefinger and holding the other fingers straight and at the same time uttering "Okay" or "Perfect". Similarly, the gesture "Come here" involves the generation of the so-called beckoning sign which in Northern America is made by sticking out and moving repeatedly the index finger from the clenched palm, facing the gesturer, and uttering a phrase such as "Come here" or "Here". We specifically address automatic detection and recognition of a set of such spontaneously generated multimodal gestures even when these are intermixed with other irrelevant actions, which could be verbal, nonverbal or both. The gesturer may, for example, be walking in-between the gestures or talking to somebody else.

In this context, we focus only on gestures that are always multimodal, that is, they are not expressed only verbally or non-verbally, without implying however strictly synchronous realizations in all modalities or making any related assumptions apart from expecting consecutive multimodal gestures to be sufficiently well separated in time, namely a few milliseconds apart in all information streams. Further, no linguistic assumptions are made regarding the sequence of gestures, namely any gesture can follow any other.

Let $V_g = \{g_i\}, i = 1, \ldots, |V_g|$ be the vocabulary of multimodal gestures g_i that are to be detected and recognized in a recording and let $S = \{\mathbf{O}_i\}, i = 1, \ldots, |S|$ be the set of information streams that are concurrently observed for that purpose. In our experiments, the latter set comprises three streams, namely audio spectral features, the gesturer's skeleton and handshape features. Based on these observations the proposed system will generate a hypothesis for the sequence of gesture appearances in a specific recording/session, like the following:

$$\mathbf{h} = [bm, g_1, sil, g_5, \ldots, bm, sil, g_3].$$

The symbol *sil* essentially corresponds to inactivity in all modalities while *bm* represents any other activity, mono- or multimodal, that does not constitute any of the target multimodal gestures. This recognized sequence is generated by exploiting single stream-based gesture models via the proposed fusion algorithm that is summarized in Figure 1 and described in detail in Section 3.2. For the sake of clarity, the single stream modeling framework is first presented in Section 3.1. Performance of the overall algorithm is found to depend on how

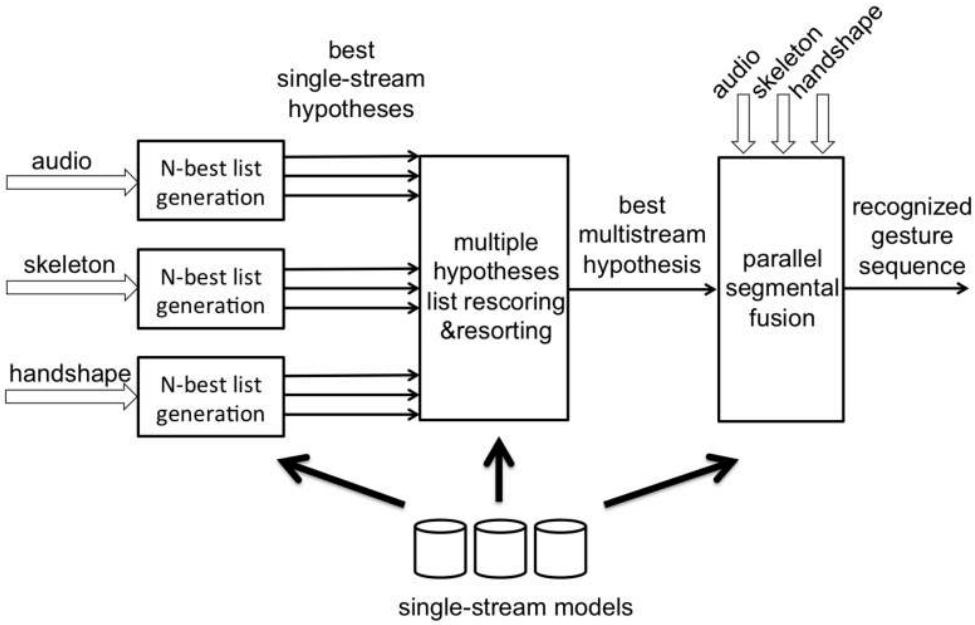


Figure 1: Overview of the proposed multimodal fusion scheme for gesture recognition based on multiple hypotheses rescoring. Single-stream models are first used to generate possible hypotheses for the observed gesture sequence. The hypotheses are then rescored by all streams and the best one is selected. Finally, the observed sequence is segmented at the temporal boundaries suggested by the selected hypothesis and parallel fusion is applied to classify the resulting segments. Details are given in Section 3.2.

accurately the single stream models represent each gesture. This representation accuracy can be significantly improved by the application of the multimodal activity detection scheme described in Section 3.3.

3.1 Speech, Skeleton and Handshape Modeling

The underlying single-stream modeling scheme is based on Hidden Markov Models (HMMs) and builds on the keyword-filler paradigm that was originally introduced for speech (Wilpon et al., 1990; Rose and Paul, 1990) in applications like spoken document indexing and retrieval (Foote, 1999) or speech surveillance (Rose, 1992). The problem of recognizing a limited number of gestures in an observed sequence comprising other heterogeneous events as well, is seen as a keyword detection problem. The gestures to be recognized are the keywords and all the rest is ignored. Then, for every information stream, each gesture $g_i \in V_g$, or, in practice, its projection on that stream, is modeled by an HMM and there are two separate filler HMMs to represent either silence/inactivity (sil) or all other possible events (bm) appearing in that stream.

All these models are basically left-to-right HMMs 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 (Young et al., 2002). Training is performed using the Baum-Welch algorithm (Rabiner and Juang, 1993), and mixture components are incrementally refined.

While this is the general training procedure followed, two alternative approaches are investigated, regarding the exact definition and the supervised training process of all involved models. These are described in the following. We experiment with both approaches and we show that increased modeling accuracy at the single-stream level leads to better results overall.

3.1.1 Training Without Employing Activity Detection

In this case, single-stream models are initialized and trained based on coarse, multimodal temporal annotations of the gestures. These annotations are common for all streams and given that there is no absolute synchronization across modalities they may also include inactivity or other irrelevant events in the beginning or end of the target gestural expression. In this way the gesture models already include, by default, inactivity segments. As a consequence we do not train any separate inactivity (*sil*) model. At the same time, the background model (*bm*) is trained on all training instances of all the gestures, capturing in this way only generic gesture properties that are expected to characterize a non-target gesture. The advantage of this approach is that it may inherently capture cross-modal synchronicity relationships. For example, the waving hand motion may start before speech in the waving gesture and so there is probably some silence (or other events) to be expected before the utterance of a multimodal gesture (e.g. "Bye bye") which is modeled implicitly.

3.1.2 Training With Activity Detection

On the other hand, training of single-stream models can be performed completely independently using stream-specific temporal boundaries of the target expressions. In this direction, we applied an activity detection scheme, described in detail in Section 3.3. Based on that, it is possible to obtain tighter stream-specific boundaries for each gesture. Gesture models are now trained using these tighter boundaries, the *sil* model is trained on segments of inactivity (different for each modality) and the *bm* model is trained on segments of activity but outside the target areas. In this case, single-stream gesture models can be more accurate but any possible evidence regarding synchronicity across modalities is lost.

3.2 Multimodal Fusion of Speech, Skeleton and Handshape

Using the single-stream gesture models (see Section 3.1) and a gesture-loop grammar as shown in Figure 2(a) we initially generate a list of N-best possible hypotheses for the unknown gesture sequence for each stream. Specifically, the Viterbi algorithm (Rabiner and Juang, 1993) is used to directly estimate the best stream-based possible hypothesis $\hat{\mathbf{h}}_m$

Algorithm 1 Multimodal Scoring and Resorting of Hypotheses

```
% N-best list rescoring
for all hypotheses do
    % Create a constrained grammar
    keep the sequence of gestures fixed
    allow insertion/deletion of sil and bm occurrences between gestures
    for all modalities do
        by applying the constrained grammar and Viterbi decoding:
        1) find the best state sequence given the observations
        2) save corresponding score and temporal boundaries
    % Late fusion to rescore hypotheses
    final hypothesis score is a weighted sum of modality-based scores
the best hypothesis of the 1st-pass is the one with the maximum score
```

for the unknown gesture sequence as follows:

$$\hat{\mathbf{h}}_m = \arg\max_{\mathbf{h}_m \in G} \log P(\mathbf{O}_m | \mathbf{h}_m, \lambda_m), \quad m = 1, \ldots, |S|$$

where $\mathbf{O}_m$ is the observation[1] sequence for modality m, λ_m is the corresponding set of models and G is the set of alternative hypotheses allowed by the gesture loop grammar. Instead of keeping just the best scoring sequence we apply essentially a variation of the Viterbi algorithm, namely the lattice N-best algorithm (Shwartz and Austin, 1991), that apart from storing just the single best gesture at each node it also records additional best-scoring gestures together with their scores. Based on these records, a list of N-best hypotheses for the entire recording and for each modality can finally be estimated.

The N-best lists are generated independently for each stream and the final superset of the multimodally generated hypotheses may contain multiple instances of the same gesture sequence. By removing possible duplicates we end up with L hypotheses forming the set $H = \{\mathbf{h}_1, \ldots, \mathbf{h}_L\}$; $\mathbf{h}_i$ is a gesture sequence (possibly including *sil* and *bm* occurrences as well). Our goal is to sort this set and identify the most likely hypothesis this time exploiting all modalities together.

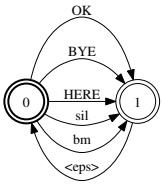
3.2.1 Multimodal Scoring and Resorting of Hypotheses

In this direction, and as summarized in Algorithm 1, we estimate a combined score for each possible gesture sequence as a weighted sum of modality-based scores

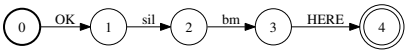
$$v_i = \sum_{m \in S} w_m v_{m,i}^s, \quad i = 1 \ldots L, \tag{1}$$

where the weights w_m are determined experimentally in a left-out validation set of multimodal recordings. The validation set is distinct from the final evaluation (test) set; more

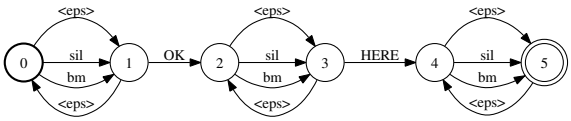
1. For the case of video data an observation corresponds to a single image frame; for the case of audio modality it corresponds to a 25 msec window.



(a) gesture-loop



(b) hypothesis



(c) hypothesis-dependent grammar

Figure 2: Finite-state-automaton (FSA) representations of finite state grammars: (a) an example gesture-loop grammar with 3 gestures plus inactivity and background labels. The "eps" transition represents an ϵ transition of the FSA, (b) an example hypothesis, (c) a hypothesis-dependent grammar allowing varying *sil* and *bm* occurrences between gestures.

Algorithm 2 Segmental Parallel Fusion

```
% Parallel scoring
for all modalities do segment observations based on given temporal boundaries
    for all resulting segments do
        estimate a score for each gesture given the segment observations
        temporally align modality segments
        for all aligned segments do
            estimate weighted sum of modality-based scores for all gestures
            select the best-scoring gesture (sil and bm included)
```

details on the selection of weights are provided in Section 5. The modality-based scores $v^s_{m,i}$ are standardized versions[2] of $v_{m,i}$ which are estimated by means of Viterbi decoding as follows:

$$v_{m,i} = \max_{\mathbf{h} \in G_{h_i}} \log P(\mathbf{O}_m | \mathbf{h}, \lambda_m), \quad i = 1, \ldots, L, \ m = 1, \ldots, |S| \tag{2}$$

where $\mathbf{O}_m$ is the observation sequence for modality m and λ_m is the corresponding set of models. This actually solves a constrained recognition problem in which acceptable gesture sequences need to follow a specific hypothesis-dependent finite state grammar G_{h_i}. It is required that the search space of possible state sequences only includes sequences

2. That is, transformed to have zero mean and a standard deviation of one.

corresponding to the hypothesis $\mathbf{h}_i$ plus possible variations by keeping the appearances of target gestures unaltered and only allow *sil* and *bm* labels to be inserted, deleted and substituted with each other. An example of a hypothesis and the corresponding grammar is shown in Figure 2(b,c). In this way, the scoring scheme accounts for inactivity or non-targeted activity that is not necessarily multimodal, e.g., the gesturer is standing still but speaking or is walking silently. This is shown to lead to additional improvements when compared to a simple forced-alignment based approach.

It should be mentioned that hypothesis scoring via (2) can be skipped for the modalities based on which the particular hypothesis was originally generated. These scores are already available from the initial N-best list estimation described earlier.

The best hypothesis at this stage is the one with the maximum combined score as estimated by (1). Together with the corresponding temporal boundaries of the included gesture occurrences, which can be different for the involved modalities, this hypothesized gesture sequence is passed on to the segmental parallel scoring stage. At this last stage, only local refinements are allowed by exploiting possible benefits of a segmental classification process.

3.2.2 Segmental Parallel Fusion

The segmental parallel fusion algorithm is summarized in Algorithm 2. Herein we exploit the modality-specific time boundaries for the most likely gesture sequence determined in the previous step, to reduce the recognition problem into a segmental classification one. First, we segment the audio, skeleton and handshape observation streams employing these boundaries. Given that in-between gestures, i.e., for *sil* or *bm* parts, there may not be one-to-one correspondence between segments of different observation streams these segments are first aligned with each other across modalities by performing an optimal symbolic string match using dynamic programming. Then, for every aligned segment t and each information stream m we compute the log probability

$$LL^t_{m,j} = \max_{\mathbf{q}\in Q} \log P(\mathbf{O}^t_m, \mathbf{q}|\lambda_{m,j}), \quad j = 1, \ldots, |V_g| + 2,$$

where $\lambda_{m,j}$ are the parameters of the model for the gesture g_j in the extended vocabulary $V_g \cup \{sil, bm\}$ and the stream $m \in S$; $\mathbf{q}$ is a possible state ($\in Q$) sequence. These segmental scores are linearly combined across modalities to get a multimodal gestural score (left hand side) for each segment

$$LL^t_j = \sum_{m\in S} w'_m LL^t_{m,j}, \tag{3}$$

where w'_m, is the stream-weight for modality m set to optimize recognition performance in a validation data set.[3] Finally, the gesture with the highest score is the recognized one for each segment t. This final stage is expected to give additional improvements and correct false alarms by seeking loosely overlapping multimodal evidence in support of each hypothesized gesture.

3. The w'_m are different from the weights in (1). Their selection is similarly based on a separate validation set that is distinct from the final evaluation set; more details on the selection of weights are provided in Section 5.

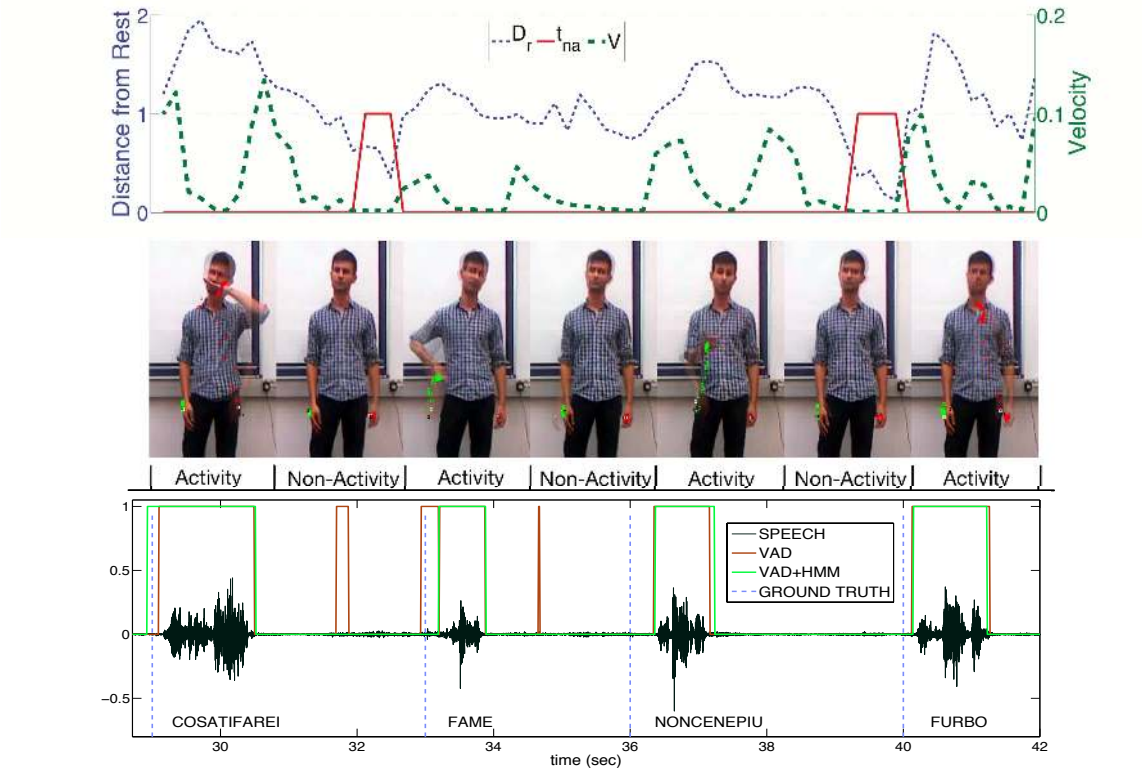


Figure 3: Activity detection example for both audio and visual modalities for one utterance. First row: The velocity of the hands (V), their distance with respect to the rest position (D_r) and the resulting initial estimation of gesture non-activity segments (t_{na}). Second row: The estimated gesture activity depicted on the actual video images. Third row: The speech signal accompanied with the initial VAD, the VAD+HMM and the gesture-level temporal boundaries included in the gesture data set (ground truth).

3.3 Multimodal Activity Detection

To achieve activity detection for each one of visual and audio modalities, we follow a common model-based framework. This is based on two complementary models of "activity" and "non-activity". In practice, these models, have different interpretations for the different modalities. This is first due to the nature of each modality, and second due to challenging data acquisition conditions. For the case of speech, the non-activity model may correspond to noisy conditions, e.g., keyboard typing or fan noise. For the case of the visual modality, the non-activity model refers to the rest cases in-between the articulation of gestures. However, these rests are not strictly defined, since the subject may not always perform a full rest and/or the hands may not stop moving. All cases of activity, in both the audio and the skeleton streams, such as out-of-vocabulary multimodal gestures and other

spontaneous gestures are thought to be represented by the activity model. Each modality's activity detector is initialized by a modality-specific front-end, as described in the following.

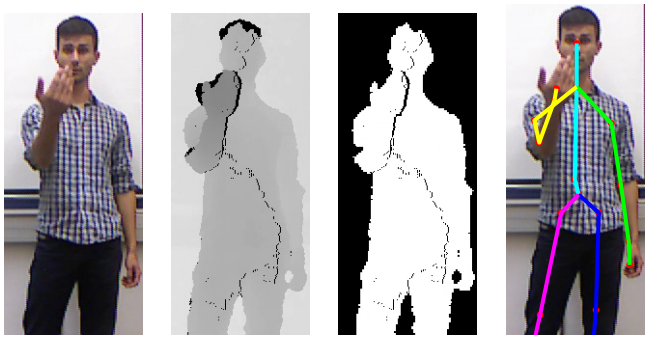
For the case of speech, activity and non-activity models are initialized on activity and non-activity segments correspondingly. These are determined by taking advantage for initialization of a Voice Activity Detection (VAD) method recently proposed by Tan et al. (2010). This method is based on likelihood ratio tests (LRTs) and by treating the LRT's for the voice/unvoiced frames differently it gives improved results than conventional LRT-based and standard VADs. The activity and non-activity HMM models are further trained using an iterative procedure employing the Baum-Welch algorithm, better known as embedded re-estimation (Young et al., 2002). The final boundaries of the speech activity and non-activity segments are determined by application of the Viterbi algorithm.

For the visual modality, the goal is to detect activity concerning the dynamic gesture movements versus the rest cases. For this purpose, we first initialize our non-activity models on rest position segments which are determined on a recording basis. For these segments skeleton movement is characterized by low velocity and the skeleton is close to the rest position $\mathbf{x_r}$. To identify non-active segments, we need to estimate a) the skeleton rest position b) the hands velocity, and c) the distance of the skeleton to that position. Hands' velocity is computed as $V(\mathbf{x}) = \|\dot{\mathbf{x}}\|$ where $\mathbf{x}(t)$ is the 3D hands' centroid coordinate vector and t is time. The rest position is estimated as the median skeleton position of all the segments for which hands' velocity V is below a certain threshold $V_{T_r} = 0.2 \cdot \bar{V}$, where $\bar{V}$ is the average velocity of all segments. The distance of the skeleton to the rest position is determined as $D_r(\mathbf{x}) = \|\mathbf{x} - \mathbf{x_r}\|$. Initial non-activity segments t_{na} are the ones for which the following two criteria hold, namely $\mathbf{t_{na}} = \{t : \ D_r(\mathbf{x}) < D_{T_r} and\ V(\mathbf{x}) < V_{T_r}\}$. Taking as input these t_{na} segments we train a non-activity HMM model while an activity model is trained on all remaining segments using the skeleton feature vector as described in Section 5.1.1. Further, similar to the case of speech we re-train the HMM models using embedded re-estimation. The final boundaries of the visual activity and non-activity segments are determined by application of the Viterbi algorithm.

In Figure 3, we illustrate an example of the activity detection for both audio and visual modalities for one utterance. In the first row, we depict the velocity of the hands (V), their distance with respect to the rest position (D_r) and the initial estimation of gesture non-activity (t_{na}) segments. We observe that in t_{na} segments both V and D_r are lower than the predefined thresholds ($V_{T_r} = 0.6, D_{T_r} = 0.006$)[4] and correspond to non-activity. In the second row, we illustrate the actual video frames images. These are marked with the tracking of both hands and accompanied with the final model-based gesture activity detection. In the bottom, we show the speech signal, with the initial VAD boundaries, the refined, HMM-based ones (VAD+HMM) and the gesture-level boundaries included in the data set (ground truth). As observed the refined detection (VAD+HMM) is tighter and more precise compared to the initial VAD and the data set annotations.

To sum up, after applying the activity detectors for both audio and visual modalities we merge the corresponding outputs with the gesture-level data set annotations in order to obtain refined stream-specific boundaries that align to the actual activities. In this way,

4. These parameters are set after experimentation in a single video of the validation set, that was annotated in terms of activity.



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

Figure 4: Sample cues of the multimodal gesture challenge 2013 data set.

we compensate for the fact that the data set annotations may contain non-activity at the start/end of each gesture.

4. Multimodal Gestures' Data set

For our experiments we employ the ChaLearn multimodal gesture challenge data set, introduced by Escalera et al. (2013b). Other similar data sets are described by Ruffieux et al. (2013, 2014). This data set focuses on multiple instance, user independent learning of gestures from multi-modal data. It provides via Kinect RGB and depth images of face and body, user masks, skeleton information, joint orientation as well as concurrently recorded audio including the speech utterance accompanying/describing the gesture (see Figure 4). The vocabulary contains 20 Italian cultural-anthropological gestures. The data set contains three separate sets, namely for development, validation and final evaluation, including 39 users and 13858 gesture-word instances in total. All instances have been manually transcribed and loosely end-pointed. The corresponding temporal boundaries are also provided; these temporal boundaries are employed during the training phase of our system.

There are several issues that render multimodal gesture recognition in this data set quite challenging as described by Escalera et al. (2013b), such as the recording of continuous sequences, the presence of distracter gestures, the relatively large number of categories, the length of the gesture sequences, and the variety of users. Further, 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 Figure 5). Similarly, single-handed gestures may be performed with either the left or right hand. Finally, variations in background, lighting and resolution, occluded body parts and spoken dialects have also been introduced.

5. Experiments

We first provide information on the multimodal statistical modeling that includes feature extraction and training. Then, we discuss the involved fusion parameters, the evaluation procedure, and finally, present results and comparisons.

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

Figure 5: (a,b) Arm position variation (low, high) for gesture 'vieni qui'; (c,d) Left and right handed instances of 'vattene'.

5.1 Parameters, Evaluation, Structure

Herein, we describe first the employed feature representations, and training parameters for each modality, such as number of states and mixture components: as discussed in Section 3.1 we statistically train separate gesture HMMs per each information stream: skeleton, handshape and audio. Next, we describe the stream weight selection procedure, note the best stream weights, and present indicative results of the procedure. After presenting the evaluation metrics, we finally describe the overall rational of the experimental structure.

5.1.1 Multimodal Features, HMM and Fusion Parameters

The features employed for the *skeleton* cue include: the hands' and elbows' 3D position, the hands 3D velocity, the 3D direction of the hands' movement, and the 3D distance of hands' centroids. For the *handshape's* representation we employ the HOG feature descriptors. These are extracted on both hands' segmented images for both RGB and depth cues. We segment the hands by performing a threshold-based depth segmentation employing the hand's tracking information. For the *audio* modality we intend to efficiently capture the spectral properties of speech signals by estimating the Mel Frequency Cepstral Coefficients (MFCCs). Our front end generates 39 acoustic features every 10 msec. Each feature vector comprises 13 MFCCs along with their first and second derivatives. All the above feature descriptors are well known in the related literature. The specific selections should not affect the conclusions as related to the main fusion contributions, since these build on the level of the likelihoods. Such an example would be the employment of other descriptors as for instance in the case of visual (e.g., Li and Allinson, 2008) or speech related features (e.g., Hermansky, 1990).

For all modalities, we train separate gesture, *sil* and *bm* models as described in Section 3.1. These models are trained either using the data set annotations or based on the input provided by the activity detectors. The number of states, Gaussian components per state, stream weights and the word insertion penalty in all modalities are determined ex-

perimentally based on the recognition performance on the validation set.[5] For skeleton, we train left-right HMMs with 12 states and 2 Gaussians per state. For handshape, the models correspondingly have 8 states and 3 Gaussians per state while speech gesture models have 22 states and 10 Gaussians per state.

The training time is on average 1 minute per skeleton and handshape model and 90 minutes per audio model. The decoding time is on average 4xRT (RT refers to real-time).[6] A significant part of the decoding time is due to the generation of the N-best lists of hypotheses. In our experiments N is chosen to be equal to 200. We further observed that the audio-based hypotheses were always ranked higher than those from the other single-stream models. This motivated us to include only these hypotheses in the set we considered for rescoring.

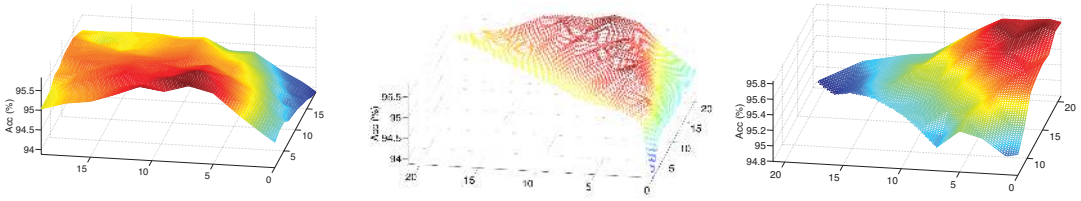
5.1.2 Stream Weight Configuration

Herein, we describe the experimental procedure for the selection of the stream weights $w_m, w'_m, m \in S$ of (1) and (3), for the components of multimodal hypothesis rescoring (MHS) and segmental parallel fusion (SPF). The final weight value selection is based on the optimization of recognition performance in the *validation* data set which is completely distinct from the final evaluation (test) data set.

Specifically, the w_m's are first selected from a set of alternative combinations to optimize gesture accuracy at the output of the MHS component. The SPF weights w'_m's are subsequently set to optimize the performance of the overall framework. The best weight combination for the multimodal hypothesis rescoring component is found to be $w^*_{SK,HS,AU} = [63.6, 9.1, 27.3]$, where SK, HS and AU correspond to skeleton, handshape and audio respectively.[7] This leads to the best possible accuracy of MHS in the validation set, namely 95.84%. Correspondingly, the best combination of weights for the segmental fusion component is $[0.6, 0.6, 98.8]$. Overall, the best achieved gesture recognition accuracy is 96.76% in the validation set.

In Figures 6(a), (b) and (c) we show the recognition accuracy of the MHS component for the various combinations of the w_m's. For visualization purposes we show accuracy when the weights vary in pairs and the remaining weight is set to its optimal value. For example, Figure 6(a) shows recognition accuracy for various combinations of handshape and audio weights when the skeleton weight is equal to 63.6. Overall, we should comment that the skeleton's contribution appears to be the most significant in the rescoring phase. This is of course a first interpretation, since the list of original hypotheses is already audio-based only, and the audio contribution cannot be directly inferred. As a consequence these results should be seen under this viewpoint. In any case, given that audio-based recognition leads to 94.1% recognition accuracy (in the validation set) it appears that both skeleton

5. Parameter ranges in the experiments for each modality are as follows. Audio: States 10-28, Gaussians: 2-32; Skeleton/handshape: States 7-15, Gaussians: 2-10.
6. For the measurements we employed an AMD Opteron(tm) Processor 6386 at 2.80GHz with 32GB RAM.
7. The weights take values in $[0, 1]$ while their sum across the modalities adds to one; these values are then scaled by 100 for the sake of numerical presentation. For the w stream weights we sampled the $[0, 1]$ with 12 samples for each modality, resulting to 1728 combinations. For the w' case, we sampled the $[0, 1]$ space by employing 5, 5 and 21 samples for the gesture, handshape and speech modalities respectively, resulting on 525 combinations.



(a) HS-AU (b) SK-AU (c) SK-HS

Figure 6: Gesture recognition accuracy of the Multiple hypothesis rescoring component for various weight-pair combinations. From left to right, the handshape-audio, skeleton-audio, skeleton-handshape weight pairs are varied. The remaining weight is set to its optimal value, namely 63.6 for skeleton, 9.1 for handshape and 27.3 for audio.

and handshape contribute in properly reranking the hypotheses and improve performance (which is again confirmed by the results in the test set presented in the following sections).

5.1.3 Evaluation

The presented evaluation metrics include the Levenshtein distance (LD)[8] which is employed in the ChaLearn publications (Escalera et al., 2013b) and the gesture recognition accuracy. The Levenshtein distance $LD(R,T)$, also known as "edit distance', is the minimum number of edit operations that one has to perform to go from symbol sequence R to T, or vice versa; edit operations include substitutions (S), insertions (I), or deletions (D). The overall score is the sum of the Levenshtein distances for all instances compared to the corresponding ground truth instances, divided by the total number of gestures. At the same time we report the standard word recognition accuracy $Acc = 1 - LD = \frac{N-S-D-I}{N}$, where N is the total number of instances of words.

Finally, we emphasize that all reported results have been generated by strictly following the original ChaLearn challenge protocol which means that they are directly comparable with the results reported by the challenge organizers and other participating teams (Escalera et al., 2013b; Wu et al., 2013; Bayer and Thierry, 2013).

5.1.4 Structure of Experiments

For the evaluation of the proposed approach we examine the following experimental aspects:

1. First, we present results on the performance of the single modality results; for these the only parameter that we switch on/off is the activity detection, which can be applied on each separate modality; see Section 5.2 and Table 1.

2. Second, we examine the performance in the multimodal cases. This main axis of experiments has as its main reference Table 2 and concerns several aspects, as follows:

   (a) Focus on the basic components of the proposed approach.

8. Note that the Levenshtein distance takes values in $[0, 1]$ and is equivalent to the word error rate.

AD	Single Modalities		
	Aud.	*Skel.*	*HS*
✗	78.4	47.6	13.3
✓	87.2	49.1	20.2

Table 1: Single modalities recognition accuracy %, including Audio (Aud.), Skeleton (Skel.), and Handshape (HS). AD refers to activity detection.

(b) Focus on two stream modality combinations; this serves for both the analysis of our approach, but also provides a more focused comparison with other methods that employ the specific pairs of modalities.

(c) Finally, we provide several fusion based variation experiments, as competitive approaches.

3. Third, we show an indicative example from the actual data, together with its decoding results after applying the proposed approach, compared to the application of a couple of subcomponents.

4. Fourth, we specifically focus on comparisons within the gesture challenge competition. From the list of 17 teams/methods that submitted their results (54 teams participated in total) we review the top-ranked ones, and list their results for comparison. Moreover, we describe the components that each of the top-ranked participants employ, providing also focused comparisons to both our complete approach, and specific cases that match the employed modalities of the other methods. Some cases of our competitive variations can be seen as resembling cases of the other teams' approaches.

5.2 Recognition Results: Single Modalities

In Table 1 we show the recognition results for each independent modality with and without the employment of activity detection (AD). Note that AD is employed for model training, as described in Sections 3.1, 3.3, for each modality. In both cases the audio appears to be the dominant modality in terms of recognition performance. For all modalities, the model-based integration of the activity detectors during training appears to be crucial: they lead to refined temporal boundaries that better align to the actual single-stream activity. In this way we compensate for the fact that the data set annotations may contain non-activity at the start/end of a gesture. By tightening these boundaries we achieve to model in more detail gesture articulation leading to more robustly trained HMMs. This is also projected on the recognition experiments: In all modalities the recognition performance increases, by 8.8%, 1.5% and 6.9% in absolute for the audio, the skeleton and the handshape streams respectively.

	Method/ Exp. Code	Modality	Segm. Method	Classifier/ Modeling	Fusion	Acc. (%)	LD
Others	O1: 1st Rank*	SK, AU	AU:time-domain	HMM, DTW	Late:w-sum	87.24	0.1280
	O2: 2nd Rank†	SK, AU	AU:energy	RF, KNN	Late:posteriors	84.61	0.1540
	O3: 3rd Rank‡	SK, AU	AU:detection	RF, Boosting	Late:w-average	82.90	0.1710
2 Streams	s2-A1	SK,AU	HMM	AD, HMM	Late:SPF	87.9	0.1210
	s2-B1	SK,AU	-	AD,HMM,GRAM	Late:MHS	92.8	0.0720
	s2-A2	HS,AU	HMM	AD, HMM	Late:SPF	87.7	0.1230
	s2-B2	HS,AU	-	AD,HMM,GRAM	Late:MHS	87.5	0.1250
3 Streams	C1	SK,AU,HS	HMM	AD, HMM	Late:SPF	88.5	0.1150
	D1	SK,AU,HS	-	HMM	Late:MHS	85.80	0.1420
	D2	SK,AU,HS	-	AD,HMM	Late:MHS	91.92	0.0808
	D3	SK,AU,HS	-	AD,HMM,GRAM	Late:MHS	93.06	0.0694
	E1	SK,AU,HS	HMM	HMM	Late:MHS+SPF	87.10	0.1290
	E2	SK,AU,HS	HMM	AD,HMM	Late:MHS+SPF	92.28	0.0772
	E3	SK,AU,HS	HMM	AD,HMM,GRAM	Late:MHS+SPF	93.33	0.0670

*(Wu et al., 2013); † (Escalera et al., 2013b); ‡ (Bayer and Thierry, 2013).

Table 2: Comparisons to first-ranked teams in the multimodal challenge recognition ChaLearn 2013, and to several variations of our approach.

5.3 Recognition Results: Multimodal Fusion

For the evaluation of the proposed fusion scheme we focus on several of its basic components. For these we refer to the experiments with codes D1-3,[9] and E1-3 as shown in Table 2. These experiments correspond to the employment of all three modalities, while altering a single component each time, wherever this makes sense.

5.3.1 Main Components and Comparisons

First comes the *MHS component* (see D1-3), which rescores the multimodal hypotheses list employing all three information streams and linearly combining their scores. Comparing with Table 1 the MHS component results in improved performance compared to the monomodal cases, by leading to 38% relative Levenshtein distance reduction (LDR)[10] on average. This improvement is statistically significant, when employing the McNemar's test (Gillick and Cox, 1989), with $p < 0.001$.[11]

Further, the employment of the activity detectors for each modality during training also affects the recognition performance after employing the MHS component, leading to a relative LDR of 38% which is statistically significant ($p < 0.001$); compare D1-D2, E1-E2.

For the N-best multimodal hypothesis rescoring we can either enforce each modality to rescore the exact hypothesis (forced alignment), or allow certain degrees of freedom by

9. The D1-3 notation refers to the D1, D2 and D3 cases.
10. All relative percentages, unless stated otherwise, refer to relative LD reduction (LDR). LDR is equivalent to the known relative word error rate reduction.
11. Statistical significance tests are computed on the raw recognition values and not on the relative improvement scores.

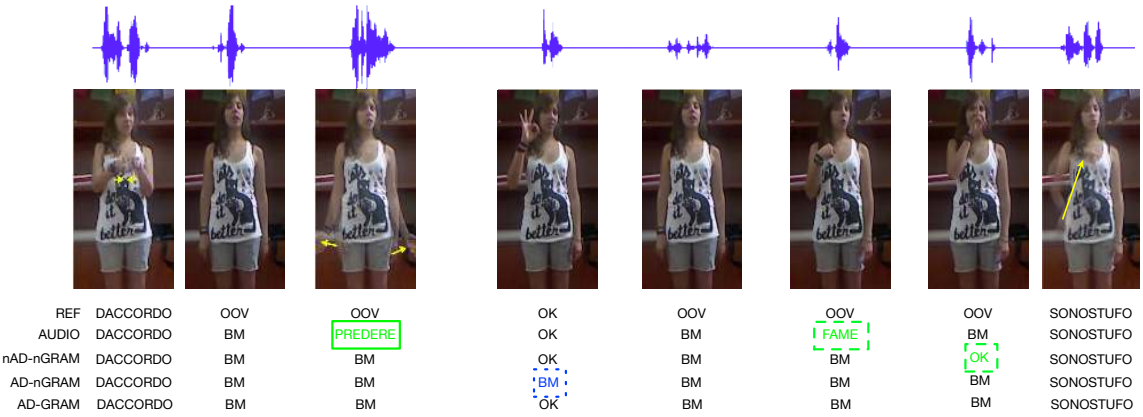


Figure 7: A gesture sequence decoding example. The audio signal is plotted in the top row the and visual modalities (second row) are illustrated via a sequence of images for a gesture sequence. Ground truth transcriptions are denoted by "REF". Decoding results are given for the single-audio modality (AUDIO) and the proposed fusion scheme employing or not the activity detection (AD) or the grammar (GRAM). In nAD-nGRAM we do not employ neither AD nor GRAM during rescoring, in AD-nGRAM we only employ AD but not GRAM and in AD-GRAM both AD and GRAM are employed. Errors are highlighted as: deletions, in blue color, and insertions in green. A background model (bm) models the out-of-vocabulary (OOV) gestures.

employing a *specific grammar* (GRAM) which allows insertions or deletions of either *bm* or *sil* models: By use of the aforementioned grammar *during rescoring* (see D2-D3, E2-E3) we get an additional 14% of relative Levenshtein distance reduction, which is statistically significant ($p < 0.001$). This is due to the fact that the specific grammar accounts for activity or non-activity that does not necessarily occur simultaneously across all different modalities.

In addition, by employing the *SPF component* (E1-3) we further refine the gesture sequence hypothesis by fusing the single-stream models at the segmental level. By comparing corresponding pairs: D1-E1, D2-E2 and D3-E3, we observe that the application of the SPF component increases the recognition performance only slightly; this increase was not found to be statistically significant. The best recognition performance, that is, 93.33%, is obtained after employing the SPF component on top of MHS, together with AD and GRAM (see E3).

On the side, we additionally provide results that account for pairs of modalities; see s2-B1 (AU+SK) and s2-B2 (AU+HS), and for the case of the *MHS component*. These two stream pair results, are comparable with the corresponding 3-stream case of D1 (plus D2-3 for additional components). The rest of the results and pairs are discussed in Section 5.4, where comparisons with other approaches are presented.

5.3.2 Example from the Results

A decoding example is shown in Figure 7. Herein we illustrate both audio and visual modalities for a word sequence accompanied with the ground truth gesture-level transcriptions (row:"REF"). In addition we show the decoding output employing the single-audio modality (AUDIO) and the proposed fusion scheme employing or not two of its basic components: activity detection (AD) and the above mentioned grammar (GRAM). In the row denoted by nAD-nGRAM we do not employ either AD or GRAM during rescoring, in the row AD-nGRAM we only employ AD but not GRAM and in AD-GRAM both AD and grammar are used. As we observe there are several cases where the subject articulates an out-of-vocabulary (OOV) gesture. 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'). When employing either the nAD-nGRAM or AD-nGRAM the above word insertions are corrected as the visual modality is integrated and helps identifying that these segments correspond to OOV gestures. Finally, both nAD-nGRAM and AD-nGRAM lead to errors which our final proposed approach manages to deal with: nAD-nGRAM causes insertion of "OK", AD-nGRAM of a word deletion "BM". On the contrary, the proposed approach recognizes the whole sentence correctly.

5.4 Comparisons

Next, we first briefly describe the main components of the top-ranked approaches in ChaLearn. This description aims at allowing for focused and fair comparisons between 1) the first-ranked approaches, and 2) variations of our approach.

5.4.1 ChaLearn First-Ranked Approaches

The first-ranked team (IV AMM) (Wu et al., 2013; Escalera et al., 2013b) uses a feature vector based on audio and skeletal information. A simple time-domain end-point detection algorithm based on joint coordinates is applied to segment continuous data sequences into candidate gesture intervals. A HMM is trained with 39-dimension MFCC features and generates confidence scores for each gesture category. A Dynamic Time Warping based skeletal feature classifier is applied to provide complementary information. The confidence scores generated by the two classifiers are firstly normalized and then combined to produce a weighted sum for late fusion. A single threshold approach is employed to classify meaningful gesture intervals from meaningless intervals caused by false detection of speech intervals.

The second-ranked team (WWEIGHT) (Escalera et al., 2013b) combines audio and skeletal information, using both joint spatial distribution and joint orientation. They first search for regions of time with high audio-energy to define time windows that potentially contained a gesture. Feature vectors are defined using a log-spaced audio spectrogram and the joint positions and orientations above the hips. At each time sample the method subtracts the average 3D position of the left and right shoulders from each 3D joint position. Data is down-sampled onto a 5Hz grid. There were 1593 features total (9 time samples x 177 features per time sample). Since some of the detected windows contain distracter gestures, an extra 21st label is introduced, defining the "not in the dictionary" gesture category. For the training of the models they employed an ensemble of randomized decision trees, referred

Rank	Approach	Lev. Dist.	Acc.%	LDR
-	Our	0.0667	93.33	-
1	iva.mm (Wu et al., 2013)	0.12756	87.244	+47.6
2	wweight	0.15387	84.613	+56.6
3	E.T. (Bayer and Thierry, 2013)	0.17105	82.895	+60.9
4	MmM	0.17215	82.785	+61.2
5	pptk	0.17325	82.675	+61.4

Table 3: Our approach in comparison with the first 5 places of the Challenge. We include recognition accuracy (Acc.) %, Levenshtein distance (Lev. Dist., see also text) and relative Levenshtein distance reduction (LDR) (equivalent to the known relative error reduction) compared to the proposed approach (Our).

to as random forests (RF), (Escalera et al., 2013b), and a k-nearest neighbor (KNN) model. The posteriors from these models are averaged with equal weight. Finally, a heuristic is used (12 gestures maximum, no repeats) to convert posteriors to a prediction for the sequence of gestures.

The third-ranked team (ET) (Bayer and Thierry, 2013; Escalera et al., 2013b) combine the output decisions of two approaches. The features considered are based on the skeleton information and the audio signal. First, they look for gesture intervals (unsupervised) using the audio and extract features from these intervals (MFCC). Using these features, they train a random forest (RF) and a gradient boosting classifier. The second approach uses simple statistics (median, var, min, max) on the first 40 frames for each gesture to build the training samples. The prediction phase uses a sliding window. The authors late fuse the two models by creating a weighted average of the outputs.

5.4.2 Comparisons With Other Approaches and Variations

Herein we compare the recognition results of our proposed multimodal recognition and multiple hypotheses fusion framework with other approaches (Escalera et al., 2013b) which have been evaluated in the exact recognition task.[12]

First, let us briefly present an overview of the results (Table 3): Among the numerous groups and approaches that participated we list the first four ones as well as the one we submitted during the challenge, that is "pptk". As shown in Table 3 the proposed approach leads to superior performance with relative LD reduction of at least 47.6%. We note that our updated approach compared to the one submitted during the challenge leads to an improvement of 61.4%, measured in terms of relative LD reduction (LDR). Compared to the approach we submitted during the challenge, the currently proposed scheme: 1) employs activity detection to train single-stream models, 2) applies the SPF on top of the MHS step, 3) introduces the grammar-constrained decoding during hypothesis rescoring and further

12. In all results presented we follow the same blind testing rules that hold in the challenge, in which we have participated (pptk team). In Table 3 we include for common reference the Levenshtein distance (LD) which was also used in the challenge results (Escalera et al., 2013b).

4) incorporates both validation and training data for the final estimation of the model parameters.

Now let us zoom into the details of the comparisons by viewing once again Table 2. In the first three rows, with side label "Others" (O1-3), we summarize the main components of each of the top-ranked approaches. These employ only the two modalities (SK+AU). The experiments with pairs of modalities s2-A1, s2-B1 can be directly compared with O1-3, since they all take advantage of the SK+AU modalities. Their differential concerns 1) the segmentation component, which is explicit for the O1-3; note that the segmentation of s2-A1 is implicit, as a by-product of the HMM recognition. 2) The modeling and recognition/classification component. 3) The fusion component. At the same time, s2-A1/s2-B1 refer to the employment of the proposed components, that is, either SPF or MHS. Specifically, s2-A1 and s2-B1 leads to at least 5% and 43.5% relative LD reduction respectively. Of course our complete system (see rest of variations) leads to even higher improvements.

Other comparisons to our proposed approach and variations are provided after comparing with the SPF-only case, by taking out the contribution of the rescoring component. In the case of all modalities, 3 stream case, (see C1) this is compared to the corresponding matching experiment E2; this (E2) only adds the MHS resulting to an improvement of 32.9% LDR. The GRAM component offers an improvement of 42% LDR (C1 vs. E3). Reduced versions compared to C1, with two-stream combinations can be found by comparing C1 with s2-A1 or s2-A2.

6. Conclusions

We have presented a complete framework for multimodal gesture recognition based on multiple hypotheses fusion, with application in automatic recognition of multimodal gestures. In this we exploit multiple cues in the visual and audio modalities, namely movement, hands' shape and speech. After employing state-of-the-art feature representations, each modality is treated under a common statistical HMM framework: this includes model-based multimodal activity detection, HMM training of gesture-words, and information fusion. Fusion is performed by generating multiple unimodal hypotheses, which after constrained rescoring and weighted combination result in the multimodally best hypothesis. Then, segmental parallel fusion across all modalities refines the final result. On the way, we employ gesture/speech background (bm) and silence (sil) models, which are initialized during the activity detection stage. This procedure allows us to train our HMMs more accurately by getting tighter temporal segmentation boundaries.

The recognition task we dealt with contains parallel gestures and spoken words, articulated freely, containing multiple sources of multimodal variability, and with on purpose false alarms. The overall framework is evaluated in a demanding multimodal data set (Escalera et al., 2013b) achieving 93.3% word accuracy. The results are compared with several approaches that participated in the related challenge (Escalera et al., 2013a), under the same blind testing conditions, leading to at least 47.6% relative Levenshtein distance reduction (equivalent to relative word error rate reduction) compared to the first-ranked team (Wu et al., 2013).

The power of the proposed fusion scheme stems from both its uniform across modalities probabilistic nature and its late character together with the multiple passes of monomodal

decoding, fusion of the hypotheses, and then parallel fusion. Apart from the experimental evidence, these features render it appealing for extensions and exploitation in multiple directions: First, the method itself can be advanced by generalizing the approach towards an iterative fusion scheme, that gives feedback back to the training/refinement stage of the statistical models. Moreover in the current generative framework, we ignore statistical dependencies across cues/modalities. These could further be examined. Second, it can be advanced by incorporating in the computational modeling specific gesture theories, e.g., from linguistics, for the gesture per se or in its multimodal version; taxonomies of gestures, e.g., that describe deictic, motor, iconic and metaphoric cases. Such varieties of cases can be systematically studied with respect to their role. This could be achieved via automatic processing of multitudes of existing data sets, which elaborate more complex speech-gesture issues, leading to valuable analysis results. Then, apart from the linguistic role of gesture, its relation to other aspects, such as, psychological, behavioral socio-cultural, or communicative, to name but a few, could further be exploited. To conclude, given the potential of the proposed approach, the acute interdisciplinary interest in multimodal gesture calls for further exploration and advancements.

Acknowledgements

This research work was supported by the project "COGNIMUSE" which is implemented under the "ARISTEIA" Action of the Operational Program Education and Lifelong Learning and is co-funded by the European Social Fund and Greek National Resources. It was also partially supported by the European Union under the project "MOBOT" with grant FP7-ICT-2011-9 2.1 - 600796. The authors want to gratefully thank Georgios Pavlakos for his contribution in previous, earlier stages, of this work.

References

U. Agris, J. Zieren, U. Canzler, B. Bauer, and K.-F. Kraiss. Recent developments in visual sign language recognition. *Universal Access in the Information Society*, 6:323–362, 2008.

J. Alon, V. Athitsos, Q. Yuan, and S. Sclaroff. A unified framework for gesture recognition and spatiotemporal gesture segmentation. *IEEE Trans. on Patt. Anal. and Mach. Intell.*, 31(9):1685–1699, 2009.

A. Argyros and M. Lourakis. Real time tracking of multiple skin-colored objects with a possibly moving camera. In *Proc. Europ. Conf. on Computer Vision*, 2004.

B. Bauer and K. F. Kraiss. Towards an automatic sign language recognition system using subunits. In *Proc. of Int'l Gest. Wrksp*, volume 2298, pages 64–75, 2001.

I. Bayer and S. Thierry. A multi modal approach to gesture recognition from audio and video data. In *Proc. of the 15th ACM Int'l Conf. on Multimodal Interaction*, pages 461–466. ACM, 2013.

P. Bernardis and M. Gentilucci. Speech and gesture share the same communication system. *Neuropsychologia*, 44(2):178–190, 2006.

N. D. Binh, E. Shuichi, and T. Ejima. Real-time hand tracking and gesture recognition system. In *Proc. of Int'l Conf. on Graphics, Vision and Image Processing (GVIP)*, pages 19–21, 2005.

A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *IEEE Trans. on Patt. Anal. and Mach. Intell.*, 23(3):257–267, 2001.

R. A. Bolt. "Put-that-there": Voice and gesture at the graphics interface. In *Proc. of the 7th Annual Conf. on Computer Graphics and Interactive Techniques*, volume 14. ACM, 1980.

H. Bourlard and S. Dupont. Subband-based speech recognition. In *Proc. Int'l Conf. on Acoustics, Speech and Sig. Proc.*, volume 2, pages 1251–1254. IEEE, 1997.

K. Bousmalis, L. Morency, and M. Pantic. Modeling hidden dynamics of multimodal cues for spontaneous agreement and disagreement recognition. In *Proc. Int'l Conf. on Autom. Face & Gest. Rec.*, pages 746–752. IEEE, 2011.

P. Buehler, M. Everingham, and A. Zisserman. Learning sign language by watching TV (using weakly aligned subtitles). In *Proc. Int'l Conf. on Comp. Vis. & Patt. Rec.*, 2009.

S. Celebi, A. S. Aydin, T. T. Temiz, and T. Arici. Gesture recognition using skeleton data with weighted dynamic time warping. *Computer Vision Theory and Applications*, 2013.

F.-S. Chen, C.-M. Fu, and C.-L. Huang. Hand gesture recognition using a real-time tracking method and hidden Markov models. *Image and Vis. Computing*, 21(8):745–758, 2003.

X. Chen and M. Koskela. Online rgb-d gesture recognition with extreme learning machines. In *Proc. of the 15th ACM Int'l Conf. on Multimodal Interaction*, pages 467–474. ACM, 2013.

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*, pages 199–202. Association for Computational Linguistics, 1989.

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.

Y. Cui and J. Weng. Appearance-based hand sign recognition from intensity image sequences. *Comp. Vis. and Im. Undrst.*, 78(2):157–176, 2000.

N. Dalal and B. Triggs. Histogram of oriented gradients for human detection. In *Proc. Int'l Conf. on Comp. Vis. & Patt. Rec.*, 2005.

W. Du and J. Piater. Hand modeling and tracking for video-based sign language recognition by robust principal component analysis. In *Proc. ECCV Wksp on Sign, Gest. and Activity*, September 2010.

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. of the 15th ACM on Int'l Conf. on Multimodal Interaction*, pages 365–368. ACM, 2013a.

S. Escalera, J. Gonzlez, X. Bar, M. Reyes, O. Lopes, I. Guyon, V. Athitsos, and H.J. Escalante. Multi-modal Gesture Recognition Challenge 2013: Dataset and Results. In *15th ACM Int'l Conf. on Multimodal Interaction (ICMI), ChaLearn Challenge and Wrksp on Multi-modal Gesture Recognition*. ACM, 2013b.

J. Foote. An overview of audio information retrieval. *Multimedia Systems*, 7(1):2–10, 1999. URL `http://link.springer.com/article/10.1007/s005300050106`.

L. Gillick and S.J. Cox. Some statistical issues in the comparison of speech recognition algorithms. In *Proc. Int'l Conf. on Acoustics, Speech and Sig. Proc.*, pages 532–535 vol.1, may 1989.

H. Glotin, D. Vergyr, C. Neti, G. Potamianos, and J. Luettin. Weighting schemes for audio-visual fusion in speech recognition. In *Proc. Int'l Conf. on Acoustics, Speech and Sig. Proc.*, volume 1, pages 173–176. IEEE, 2001.

B. Habets, S. Kita, Z. Shao, A. Özyurek, and P. Hagoort. The role of synchrony and ambiguity in speech–gesture integration during comprehension. *Journal of Cognitive Neuroscience*, 23(8):1845–1854, 2011.

J. Han, G. Awad, and A. Sutherland. Modelling and segmenting subunits for sign language recognition based on hand motion analysis. *Patt. Rec. Letters*, 30:623–633, 2009.

H. Hermansky. Perceptual linear predictive (plp) analysis of speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990.

A. Hernández-Vela, M. Á. Bautista, X. Perez-Sala, V. Ponce-López, S. Escalera, X. Baró, O. Pujol, and C. Angulo. Probability-based dynamic time warping and bag-of-visual-and-depth-words for human gesture recognition in rgb-d. *Patt. Rec. Letters*, 2013.

C.-L. Huang and S.-H. Jeng. A model-based hand gesture recognition system. *Machine Vision and Application*, 12(5):243–258, 2001.

M. Isard and A. Blake. Condensation-conditional density propagation for visual tracking. *Int'l Journal of Computer Vision*, 29(1):5–28, 1998.

M. Iverson, J. and S. Goldin-Meadow. Why people gesture when they speak. *Nature*, 396 (6708):228–228, 1998.

A. Jaimes and N. Sebe. Multimodal human–computer interaction: A survey. *Comp. Vis. and Im. Undrst.*, 108(1):116–134, 2007.

S. D. Kelly, A. Özyürek, and E. Maris. Two sides of the same coin speech and gesture mutually interact to enhance comprehension. *Psychological Science*, 21(2):260–267, 2010.

A. Kendon. *Gesture: Visible Action as Utterance*. Cambridge University Press, 2004.

W. Kong and S. Ranganath. Sign language phoneme transcription with rule-based hand trajectory segmentation. *J. Signal Proc. Sys.*, 59:211–222, 2010.

I. Laptev, M. Marszalek, and B. Schmid, C.and Rozenfeld. Learning realistic human actions from movies. In *Proc. Int'l Conf. on Comp. Vis. & Patt. Rec.*, pages 1–8. IEEE, 2008.

H-K. Lee and J-H. Kim. An HMM-based threshold model approach for gesture recognition. *IEEE Trans. on Patt. Anal. and Mach. Intell.*, 21(10):961–973, 1999.

J. Li and N. M. Allinson. Simultaneous gesture segmentation and recognition based on forward spotting accumulative hmms. *Pattern Recognition*, 40(11):3012–3026, 2007.

J. Li and N. M. Allinson. A comprehensive review of current local features for computer vision. *Neurocomputing*, 71(10):1771–1787, 2008.

D. G. Lowe. Object recognition from local scale-invariant features. In *Proc. Int'l Conf. on Comp. Vis.*, pages 1150–1157, 1999.

P. Maragos, P. Gros, A. Katsamanis, and Papandreou G. Cross-modal integration for performance improving in multimedia: A review. In P. Maragos, A. Potamianos, and P. Gros, editors, *Multimodal Processing and Interaction: Audio, Video, Text*, chapter 1, pages 3–48. Springer-Verlag, New York, 2008.

D. McNeill. *Hand and Mind: What Gestures Reveal About Thought*. University of Chicago Press, 1992.

M. Miki, N. Kitaoka, C. Miyajima, T. Nishino, and K. Takeda. Improvement of multimodal gesture and speech recognition performance using time intervals between gestures and accompanying speech. *EURASIP Journal on Audio, Speech, and Music Processing*, 2014 (1):17, 2014. URL `http://link.springer.com/article/10.1186/1687-4722-2014-2`.

D. Morris, P. Collett, P. Marsh, and O'Shaughnessy M. *Gestures: Their Origins and Distribution*. Stein and Day, 1979.

Y. Nam and K. Wohn. Recognition of space-time hand-gestures using hidden Markov model. In *ACM Symposium on Virtual Reality Software and Technology*, pages 51–58, 1996.

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. of the 15th ACM Int'l Conf. on Multimodal Interaction*, pages 475–482. ACM, 2013.

N. Neverova, C. Wolf, G. Paci, G. Sommavilla, G. Taylor, and F. Nebout. A multi-scale approach to gesture detection and recognition. In *Proc. of the IEEE Int'l Conf. on Computer Vision Wrksp*, pages 484–491, 2013.

E.-J. Ong and R. Bowden. A boosted classifier tree for hand shape detection. In *Proc. Int'l Conf. on Autom. Face & Gest. Rec.*, pages 889–894. IEEE, 2004.

M. Ostendorf, A. Kannan, S. Austin, O. Kimball, R. M. Schwartz, and J. R. Rohlicek. Integration of Diverse Recognition Methodologies Through Reevaluation of N-Best Sentence Hypotheses. In *HLT*, 1991.

S. Oviatt and P. Cohen. Perceptual user interfaces: multimodal interfaces that process what comes naturally. *Communications of the ACM*, 43(3):45–53, 2000.

G. Papandreou, A. Katsamanis, V. Pitsikalis, and P. Maragos. Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(3):423–435, 2009.

V. Pitsikalis, S. Theodorakis, C. Vogler, and P. Maragos. Advances in phonetics-based sub-unit modeling for transcription alignment and sign language recognition. In *IEEE CVPR Wksp on Gest. Rec.*, 2011.

I. Poddar, Y. Sethi, E. Ozyildiz, and R. Sharma. Toward natural gesture/speech HCI: A case study of weather narration. In *Proc. Wrksp on Perceptual User Interfaces*, 1998.

V. Ponce-López, S. Escalera, and X. Baró. Multi-modal social signal analysis for predicting agreement in conversation settings. In *Proc. of the 15th ACM Int'l Conf. on Multimodal Interaction*, pages 495–502. ACM, 2013.

G. Potamianos, C. Neti, J. Luettin, and I. Matthews. Audio-visual automatic speech recognition: An overview. *Issues in Visual and Audio-Visual Speech Processing*, 22:23, 2004.

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

Z. Ren, J. Yuan, and Z. Zhang. Robust hand gesture recognition based on finger-earth mover's distance with a commodity depth camera. In *Proc. of the 19th ACM Int'l Conf. on Multimedia*, pages 1093–1096. ACM, 2011.

R. C. Rose. Discriminant wordspotting techniques for rejecting non-vocabulary utterances in unconstrained speech. In *Proc. Int'l Conf. on Acoustics, Speech and Sig. Proc.*, volume 2, pages 105–108. IEEE, 1992. URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=226109`.

R. C. Rose and D. B. Paul. A hidden Markov model based keyword recognition system. In *Proc. Int'l Conf. on Acoustics, Speech and Sig. Proc.*, pages 129–132, 1990. URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=115555`.

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*, 14(1):1627–1663, 2013.

S. Ruffieux, D. Lalanne, and E. Mugellini. ChAirGest: A Challenge for Multimodal Mid-air Gesture Recognition for Close HCI. In *Proc. of the 15th ACM Int'l Conf. on Multimodal Interaction*, ICMI '13, pages 483–488, New York, NY, USA, 2013. ACM.

S. Ruffieux, D. Lalanne, E. Mugellini, and O. A. Khaled. A Survey of Datasets for Human Gesture Recognition. In *Human-Computer Interaction. Advanced Interaction Modalities and Techniques*, pages 337–348. Springer, 2014.

R. Sharma, M. Yeasin, N. Krahnstoever, I. Rauschert, G. Cai, I. Brewer, A. M MacEachren, and K. Sengupta. Speech-gesture driven multimodal interfaces for crisis management. *Proc. of the IEEE*, 91(9):1327–1354, 2003.

S. Shimojo and L. Shams. Sensory modalities are not separate modalities: plasticity and interactions. *Current Opinion in Neurobiology*, 11(4):505–509, 2001.

J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 56(1):116–124, 2013.

R. Shwartz and S. Austin. A comparison of several approximate algorithms for finding multiple N-Best sentence hypotheses. In *Proc. Int'l Conf. on Acoustics, Speech and Sig. Proc.*, 1991.

Y. C. Song, H. Kautz, J. Allen, M. Swift, Y. Li, J. Luo, and C. Zhang. A Markov logic framework for recognizing complex events from multimodal data. In *Proc. of the 15th ACM Int'l Conf. on Multimodal Interaction*, pages 141–148. ACM, 2013.

T. Starner, J. Weaver, and A. Pentland. Real-time american sign language recognition using desk and wearable computer based video. *IEEE Trans. on Patt. Anal. and Mach. Intell.*, 20(12):1371–1375, Dec. 1998.

L. N. Tan, B. J. Borgstrom, and A. Alwan. Voice activity detection using harmonic frequency components in likelihood ratio test. In *Proc. Int'l Conf. on Acoustics, Speech and Sig. Proc.*, pages 4466–4469. IEEE, 2010.

N. Tanibata, N. Shimada, and Y. Shirai. Extraction of hand features for recognition of sign language words. In *Proc. Int'l Conf. on Vision Interface*, pages 391–398, 2002.

S. Theodorakis, V. Pitsikalis, and P. Maragos. Dynamic-Static Unsupervised Sequentiality, Statistical Subunits and Lexicon for Sign Language Recognition. *Imave and Vision Computing*, 32(8):533549, 2014.

M. Turk. Multimodal interaction: A review. *Patt. Rec. Letters*, 36:189–195, 2014.

C. Vogler and D. Metaxas. A framework for recognizing the simultaneous aspects of american sign language. *Comp. Vis. and Im. Undrst.*, 81:358, 2001.

S. B. Wang, A. Quattoni, L. Morency, D. Demirdjian, and T. Darrell. Hidden conditional random fields for gesture recognition. In *Proc. Int'l Conf. on Comp. Vis. & Patt. Rec.*, volume 2, pages 1521–1527. IEEE, 2006.

D. Weimer and S. Ganapathy. A synthetic visual environment with hand gesturing and voice input. In *ACM SIGCHI Bulletin*, volume 20, pages 235–240. ACM, 1989.

L. D Wilcox and M. Bush. Training and search algorithms for an interactive wordspotting system. In *Proc. Int'l Conf. on Acoustics, Speech and Sig. Proc.*, volume 2, pages 97–100. IEEE, 1992.

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*, 38(11):1870–1878, 1990.

A. Wilson and A. Bobick. Parametric hidden Markov models for gesture recognition. *IEEE Trans. on Patt. Anal. and Mach. Intell.*, 21:884–900, 1999.

J. Wu, J. Cheng, C. Zhao, and H. Lu. Fusing multi-modal features for gesture recognition. In *Proc. of the 15th ACM Int'l Conf. on Multimodal Interaction*, pages 453–460. ACM, 2013.

M.-H. Yang, N. Ahuja, and M. Tabb. Extraction of 2d motion trajectories and its application to hand gesture recognition. *IEEE Trans. on Patt. Anal. and Mach. Intell.*, 24(8):1061–1074, Aug. 2002.

S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. *The HTK Book*. Entropic Cambridge Research Laboratory, Cambridge, United Kingdom, 2002.