Cite this chapter as:
A. Katsamanis, V. Pitsikalis, S. Theodorakis, and P. Maragos, "Multimodal Gesture Recognition", *The Handbook of Multimodal-Multisensor Interfaces, Volume 1: Foundations, User Modeling, and Common Modality Combinations*, S. Oviatt, B. Schuller, P. Cohen, D. Sonntag, G. Potamianos, and A. Kruger (Eds.), Ch. 11, pp. 449–487, ACM Books / Morgan-Claypool Publishers, San Rafael, CA, 2017.

11 Multimodal Gesture Recognition

Athanasios Katsamanis, Vassilis Pitsikalis,
Stavros Theodorakis, Petros Maragos

11.1 Introduction

Starting from the famous "Put That There!" demonstration prototype, developed by the Architecture Machine Group at MIT in the late 70s, the growing potential of multimodal gesture interfaces in natural human-machine communication setups has stimulated people's imagination and motivated significant research efforts in the fields of computer vision, speech recognition, multimodal sensing, fusion, and human-computer interaction (HCI). In the words of Bolt [1980, p. 1]: "Because voice can be augmented with simultaneous pointing, the free usage of pronouns becomes possible, with a corresponding gain in naturalness and economy of expression. Conversely, gesture aided by voice gains precision in its power to reference".

Multimodal gesture recognition lies at the heart of such interfaces. As also defined in the Glossary, the term refers to the complex computational task comprising three main modules: (a) tracking of human movements, primarily of the hands and arms, and recognition of characteristic such motion patterns; (b) detection of accompanying speech activity and recognition of what is spoken; and (c) combination of the available audio-visual information streams to identify the multimodally communicated message.

To successfully perform such tasks, the original "Put That There!" system of Bolt [1980] imposed certain limitations on the interaction. Specifically, it required that the user be tethered by wearing a position sensing device on the wrist to capture gesturing and a headset microphone to record speech, and it allowed multimodal manipulation via speech and gestures of a small only set of shapes on a rather large screen (see also Figure 11.1). Since then however, research efforts in the field of multimodal gesture recognition have moved beyond such limited scenarios, capturing and processing the multimodal data streams by employing distant audio and visual sensors that are unobtrusive to humans. In particular, in recent years, the introduction of affordable and compact multimodal sensors like the Microsoft ***Kinect*** has enabled robust capturing of human activity. This is due to the wealth of raw and metadata streams provided by the device, in addition to the traditional planar RGB video, such as depth scene information, multiple audio channels, and human skeleton and facial tracking, among oth-

Figure 11.1 **Screenshot from the "Put That There!" demonstration video by the Architecture Machine Group at MIT [Bolt 1980].**

ers [Kinect 2016]. Such advancements have led to intensified efforts to integrate multimodal gesture interfaces in real-life applications.

Indeed, the field of multimodal gesture recognition has been attracting increasing interest, being driven by novel HCI paradigms on a continuously expanding range of devices equipped with multimodal sensors and ever-increasing computational power, for example smartphones and smart television sets. Nevertheless, the capabilities of modern multimodal gesture systems remain limited. In particular, the set of gestures accounted for in typical setups is mostly constrained to pointing gestures, a number of emblematic ones like an open palm, and gestures corresponding to some sort of interaction with a physical object, e.g., pinching for zooming. At the same time, fusion with speech remains in most cases just an experimental feature. When compared to the abundance and variety of gestures and their interaction with speech in natural human communication, it clearly seems that there is still a long way to go for the corresponding HCI research and development [Kopp 2013].

Multimodal gesture recognition constitutes a wide multi-disciplinary field. This chapter makes an effort to provide a comprehensive overview of it, both in theoretical and application terms. More specifically, basic concepts related to gesturing, the multifaceted interplay of gestures and speech, and the importance of gestures in HCI are discussed in Section 11.2. An overview of the current trends in the field of multimodal gesture recognition is provided in Section 11.3, separately focusing on gestures, speech, and multimodal fusion. Further, a state-of-the-art recognition setup developed by the authors is described in detail in Section 11.4, in order to facilitate a better understanding of all practical considerations involved in such a system. In closing, the future of multimodal gesture recognition and related challenges are discussed in Section 11.5. Finally, a set of Focus Questions to aid comprehension of the material is also provided.

Glossary Terminology for Understanding Multimodal Gesture Recognition

Co-speech gestures are gestures produced while speaking. Their interplay with speech and their role in human interaction are discussed in Section 11.2.2.

Gesture recognition models are primarily statistical constructs that can be trained to represent specific gestures based on corresponding data. In this chapter, traditional hidden Markov models (HMMs) with Gaussian mixture model (GMM) observation probabilities are considered, as part of the authors' gesture recognition system detailed in Section 11.4. Further, a number of models based on deep learning approaches are overviewed in Section 11.3.1. A more elaborate discussion of deep learning for multimodal interaction modeling can be found in the second volume of this Handbook (e.g., [Keren et al. 2017]).

Gesture types: Following the classification scheme of McNeill [1992], there exist five basic gesture types, namely **emblems**, **icons**, **metaphors**, **deictics**, and **beats**. Their main properties and typical examples are discussed in Section 11.2.1, as well as in Chapter 6 of this volume [Kopp and Bergmann 2017].

Kinect is a multimodal sensing device containing an array of specialized cameras and microphones to facilitate advanced audio-visual scene understanding [Kinect 2016]. It was originally developed as an accessory for a gaming machine, but has since found numerous applications in HCI research. The Kinect provides multiple raw data streams, namely traditional planar RGB video, video frames with depth information of the scene, and five audio channels (one for each of its four microphones, as well as a single enhanced audio signal). It is therefore often referred to as an **RGB-D-A** sensor, or simply as an **RGB-D** one, if only the video data are considered. Further, the Kinect provides metadata streams, for example detailed 3D position, pose, and skeleton joint coordinates of human subjects in the scene, a facial mesh for nearby subjects in relatively frontal head pose, detection of a small set of hand gestures, etc.

Multimodal gesture recognition refers in this chapter to the composite computational task of identifying characteristic motion patterns of hands and arms (primarily), recognizing accompanying speech, and, finally, combining the available audio-visual information streams to identify the multimodally communicated message by means of the hand and arm movements and the uttered speech.

11.2 Multimodal Communication and Gestures

According to Kendon [2004, p. 1], a gesture is "a visible action as utterance", or, in other words, "a [visible] action that is employed as a part of the process of discourse, as part of

uttering something to another in an explicit manner". People, for example, point to something to refer to it, or move their hands and arms appropriately to show the size of an object or illustrate an abstract idea. In general, gestures correspond to coordinated motion involving at least some of the following: hands, arms, torso, and head; in most cases though, the research community focuses on gestures formed by hand and arm movements.

Gesturing is practically ubiquitous and constitutes an integral part of human communication that is inherently multimodal. This chapter will mainly focus on ***co-speech gestures***, namely gesturing performed while speaking. There is significant evidence that such gesturing can be viewed as yet another aspect of speech as a whole, like, for example, its emotional and syntactic aspects [McNeill 1985, McNeill et al. 2008], being pragmatically co-expressive with them and not redundant. Further, like emotions or syntax in speech, gesturing is typically taken for granted by interlocutors. Almost 90% of spoken utterances in descriptive conversations are accompanied by speech-synchronized gestures [Nobe 2000], which occur in similar form for speakers across many cultures, independently of the language they speak [McNeill et al. 2008]. Gesturing is to a great extent unconscious, but still constitutes an important component of speech and language production.

In the rest of the section, we briefly present the fundamental properties of gestures and, most significantly, discuss the multifaceted speech-gesture interplay and its importance in multimodal human communication. The role of gestures in HCI and relevant application examples are also reviewed.

11.2.1 Gesture Types and Phases

Various researchers have tried to categorize gestures along a number of relevant dimensions, but it seems that a widely accepted typology covering the multiplicity of gestures in use by speakers globally is yet to be proposed. According to Kendon [2004, p. 107], "a single unified classification scheme of gesture is merely impossible given the multitude of dimensions gesture can depend on". In the following, we briefly present the basic ***gesture types*** (also reviewed by Kopp and Bergmann [2017]), based on the classification scheme proposed by McNeill [1992] with respect to the semantics of gestures. For brevity, this presentation is by no means exhaustive, but provides a number of illustratory examples of typical multimodal gestures. For a more comprehensive account of gesture typologies, the reader is referred to [Müller et al. 2013].

- ***Emblems***: Emblematic gestures convey a particular meaning that is widely accepted by people of a specific culture [Efron 1941, Ekman and Friesen 1969], e.g., the "thumbs up" gesture, which means "good" in North America, or the "V" gesture signifying "victory". An example of emblematic gesture "OK" is depicted in Figure 11.2. These gestures are typically stand-alone and do not require any accompanying speech to become understood.

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

Figure 11.2 **An example realization of the emblematic gesture "OK", depicting video frames of all five phases of its formation:** (a) initial phase; (b) preparation; (c) pre-stroke hold; (d) stroke; and (e) retraction. The pre-stroke hold phase is exactly when the gesturer has raised his hand and pauses instantly before the stroke, i.e., the abrupt forward movement of the hand.

- ***Icons***: Iconic gestures are motion patterns conveying a physical property of a described object, action, or event. For example, they may indicate the shape or size of an object, as in bringing the thumb and index together to form a circle describing a camera lens. Iconic gestures may also show how a certain action is executed, independently of whether this information is in the accompanying speech or not. For example, one may refer to a screwdriver and how unscrewing is performed, by repeatedly opening, closing, and rotating one's fist, while at the same time saying "It took me a while to remove that screw".
- ***Metaphors***: Metaphoric gestures are similar to iconic gestures, in that they also represent a certain property of the described concept, however, in this case, the property or concept are abstract. An example is moving the palm in a way that alludes to rolling motion, accompanying the phrase "the talk went on and on".
- ***Deictics***: Deictic gestures localize, namely indicate a certain location in space. Such could be either in the physical space in front of the speaker or the listener, or in a conceptual space defined by the interlocutors during their conversation. For example, one may point downwards to refer to "here and now", or point into the distance to refer to somebody who is absent or to an event in the future.
- ***Beats***: Beat gestures are simple, rapid, and repetitive movements, typically synchronous with prosodic variations of speech. They are used to emphasize certain parts of what is being said. An example is when the speaker emphatically says "we need to act now", accompanied by his hand flicking up and down following the rhythm of his speech.

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

Figure 11.3 **Examples of two gestures from the ChaLearn corpus [Escalera et al. 2013b], each shown in two different realizations:** (a, b) Gesture "vieni qui" ("come here"), performed by two different gesturers exhibiting significant variation in arm placement; (c, d) gesture "vattene" ("go away"), performed by the same gesturer using the left vs. the right hand. In all examples, motion is shown by superimposing the start- and end-frames of the gesture, with arrows highlighting movement direction. In terms of typology, both gestures can be attributed with deixis and iconicity, as discussed in Section 11.2.1. (From Maragos et al. [2016])

It is important to note that the aforementioned typology should only be considered as defining a number of dimensions, along which a gesture can be represented (e.g., iconicity, deixis, metaphoricity), and it does not really provide disjoint categories [McNeill 2005]. For example, a dynamic pointing gesture that also shows the direction of a movement should be attributed with both iconicity and deixis [Wagner et al. 2014]. Examples of two such gestures are depicted in Figure 11.3. There, two realizations are shown for each, demonstrating the inherent variability in their formation.

Focusing on gestures performed by a coordinated motion of hands and arms, it is also of interest to briefly review the common properties of their dynamics. More specifically, the following five phases have been identified in gesture formation [Bressem and Ladewig 2011, Wagner et al. 2014]:

- The initial phase, where the body is at a fixed rest position, and the gesticulation starts from there;
- a preparation phase, in which the hands and arms start to move away from the rest position in preparation for the following phase;
- a gesture stroke, which is the most important phase of the gesture in terms of communicating the intended message, and typically includes a peak in effort and informativeness;
- holds, which are static phases potentially preceding or following the stroke; and, finally,

- a retraction or recovery phase, when the hands and arms move back to a rest position.

A sequence of snapshots of these phases during a realization of the emblematic gesture "OK" is shown in Figure 11.2.

11.2.2 Forms of Speech-Gesture Interplay

As discussed earlier, gestures and speech are strongly interconnected. Indeed, there are numerous studies reporting and analyzing the interrelations between gestures and spoken language [Goldin-Meadow and Alibali 2013] (for cognitive models of these interrelations see also [Kopp and Bergmann 2017]). In the following, we present the three main forms of speech-gesture interplay, namely in information communication, speech production, and language learning.

Role of Gestures in Communication

The communicative role of gestures had been a controversial topic for over 35 years [Krauss 1998], with a relative consensus on it only recently reached. Based on a growing body of evidence, it is nowadays widely accepted that gestures significantly aid human communication, by often elaborating upon and enhancing the information that is orally co-expressed [Hostetter 2011, McNeill 1992]. One may, for example, use hand motions to describe the shape of an object, while only referring to its large size by speech. Further, in noisy situations, or when speech is ambiguous [Thompson and Massaro 1986], it has been found that listeners pay attention to gestures to make sense of what is being said. This behavior becomes even more prominent as the audio signal-to-noise ratio decreases. Gestures also contribute to a more accurate understanding of narratives, even in cases when speech is comprehensible [Singer and Goldin-Meadow 2005]. Interestingly, it seems that synchronously presented and non-redundant speech and gestures, independently of whether they are congruent or incongruent, are treated equivalently by listeners while trying to create a single multimodal mental representation of the available information [Cassell et al. 1999]. For example, when adults are asked to evaluate how a child performs, e.g., in a problem-solving task, they build their assessment also based on information conveyed via the child's gestures, even when such may not be in agreement with what the child said [Goldin-Meadow et al. 1992].

Another illustrative phenomenon related to the communicative nature of gestures and its relation with speech is that speakers rarely make gesture errors. Speech disfluencies may be quite common, but, in contrast, gestures have been observed to almost always convey what the speaker truly wants to communicate [Cassell 1998]. The speaker most probably will point towards the left when he means "left", although he may mistakenly say "right". This is exploited by listeners, who may be correcting speech errors based on the co-occurring gestures [McNeill 1992].

Role of Gestures in Speech Production

Gestures are also important in speech production. According to Krauss [1998], this is because they facilitate lexical access, or, based on Goldin-Meadow et al. [2001], because they lighten speaker cognitive load during speech production, by enriching the way information has been encoded and can be accessed, or by helping speakers organize it better. It is indeed true that gestures are performed even when they cannot be seen by listeners. For example, congenitally blind speakers gesture when speaking to blind listeners [Iverson and Goldin-Meadow 1998], and people gesture when talking on the phone [Rimé 1982]. So, gestures must serve an additional role for the speaker, apart from just enhancing face-to-face communication.

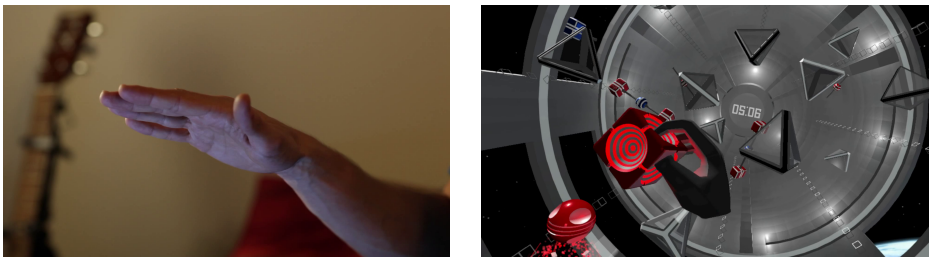
The "Lexical Retrieval Facilitation" hypothesis has been endorsed to provide an explanation for such observations [de Ruiter 2006, Krauss 1998]. The hypothesis is supported by numerous studies [Krauss et al. 2000], and it starts from the premise that knowledge is stored in memory in various domains/formats, e.g., visuo-spatial and motoric. It then states that gesturing stimulates additional memory representations, e.g., spatio-dynamic, thus cross-modally priming lexical items. For example, when a speaker performs a circular gesture as he says "the ball rolled down the hill", this gesture will contribute to the activation of word "roll", thus making it more easily accessible.

To further support this, studies have shown that, when lexical access becomes harder, speakers gesture at higher rates [Morsella and Krauss 2004], and, when speakers are not allowed to gesture, they become more disfluent. It has also been found that the duration of a gesture is related to how long it takes for a speaker to retrieve a word [Morrel-Samuels and Krauss 1992]. Further, more gestures occur during spontaneous speech than in rehearsed, apparently due to the stronger demand for on-the-fly word retrieval [Chawla and Krauss 1994]. In addition, when gesturing is constrained, it seems that people speak more slowly if speech semantics are related to space [Rauscher et al. 1996]. Along the same lines, children were found to perform worse in a naming task in the gesturing-prohibited vs. the gesturing-allowed condition [Pine et al. 2007].

Additional evidence for the importance of gestures in speech production comes from medical studies. For example, stroke patients with problems in word naming and lexical retrieval, but with satisfactory comprehension, were found to gesture more during narration than typical patients or others with visuo-spatial deficits [Hadar et al. 1998a]. Further, patients suffering from Wernicke's aphasia, i.e., having problems in comprehension but not in object naming, were found to not gesture as much [Hadar et al. 1998b]. Last, aphasic patients with phonological access or storage deficits, participating in a naming task, were significantly aided when told to point, visualize, and gesture [Rose and Douglas 2001].

Role of Gestures in Language Learning

Finally, gesturing has also been found to be important for language development in children. In particular, it has been found that the use of iconic gestures by mothers can reliably



(a) (b)

Figure 11.4 **Two examples of HCI systems employing gestures:** (a) A gesture interface controlling computer functionality (https://www.youtube.com/watch?v=Wq1FM84uAck); (b) gesture recognition in virtual reality (https://www.youtube.com/watch?v=ALBsLEupolY).

predict noun acquisition and comprehension by children [Zammit and Schafer 2011]. Being synchronized with speech, gestures seem to aid infants deduce the meaning of the words they hear [Zukow-Goldring 1996]. Interestingly, most infants typically start using their own gestures to communicate with parents before the age of ten months, namely significantly earlier than when they start to speak. The gestures initially used by children are deictic, e.g., pointing, or representational, i.e., iconic and abstract deictic [McNeill 1992], but they become more complex as children grow older. There is, it seems, a strong correlation between age, narrative complexity, and gesturing [Colletta et al. 2010]. Co-speech gestures apparently develop in parallel with linguistic capabilities.

11.2.3 Gestures in Human-Computer Interaction

The gestures mostly considered in computer vision and HCI systems are the emblematic ones, e.g., putting the hand up to mean "stop", as well as deictic ones, when referring to objects or locations in the physical space, e.g., pointing at a book and then a table while saying "put that there". The emblematic gestures are those that come to mind as prototypical or exemplary ones, but they typically constitute less than 10% of all gestures produced. However, they are used a lot in HCI systems, because they are easier for people to remember, and their realizations are relatively less variant across performers. Such gestures are typically employed as commands to control the functionality of a computer, tablet, or a smart television set, e.g., to browse the music library, play the next song, raise the volume, or play the selected movie (see Figure 11.4a).

Deictic gestures are also accounted for in HCI, because they can be very useful when one wants to perform a certain task, as discussed by Bolt [1980]. These are often used in cases when the intended communication is related to the space around the speaker, physical or virtual. Examples include controlling a virtual control panel by pushing buttons, pulling

sliders (see Figure 11.4b), or browsing through and selecting movies in a TV set movie library [Shah 2012].

The majority however of co-speech gestures, despite their importance in communication as discussed earlier, is still largely ignored by HCI systems [Cassell 1998]. This observation was first made almost 20 years ago, but things have not changed much in this respect since. A computer interface that would benefit from all the cues humans use in natural interaction is yet to be developed. The variability of everyday gestures, which are mostly produced unconsciously, and their possibly indirect connection to the content of accompanying speech render their recognition quite challenging. Making sense of them independently of speech is almost impossible, as it can also be very hard to fully understand speech without them. That is, for example, the case when a user tries to describe the relative size of an object by saying "it is this big", while protruding both palms and holding them at a certain distance from each other. In this context, naturalistic HCI systems have to first generate an appropriate multimodal semantic representation of the user state and then react accordingly. This is why the need for multimodal gesture recognition becomes even more acute.

11.3 Recognizing Speech and Gestures

A multimodal gesture recognition system will typically comprise a module for classifying hand and arm gestures, a speech recognition component, and a module to fuse results based on the available modalities, as shown in Figure 11.5. The goal of such a system is to reliably determine what the speaker has multimodally communicated.

11.3.1 Gesture Recognition

Gesture recognition has been in the focus of intense research activity for over 40 years [Mitra and Acharya 2007, Pavlovic et al. 1997]. A comprehensive overview of all related literature is beyond the scope of this chapter, and the interested reader is referred to a number of excellent reviews of the field [Chaudhary et al. 2011, Pisharady and Saerbeck 2015, Rautaray and Agrawal 2015]. Instead, this section will focus on describing the most recently proposed approaches that constitute the current trends concerning ***gesture recognition models***. Such are mainly based on deep learning techniques, a topic thoroughly reviewed in the second volume of this Handbook (e.g., [Keren et al. 2017]). Similarly to other domains where machine learning plays a central role [Deng and Yu 2014], deep learning methods provide significant modeling power, taking advantage of large corpora available for training and recent strides in parallel computing. Such techniques have the potential to lead to even greater performance improvements in the future.

Statistical Gesture Modeling

Probably the most significant aspect of the majority of gestures and, more generally, human actions that needs to be accounted for in a recognition system is their dynamic nature.

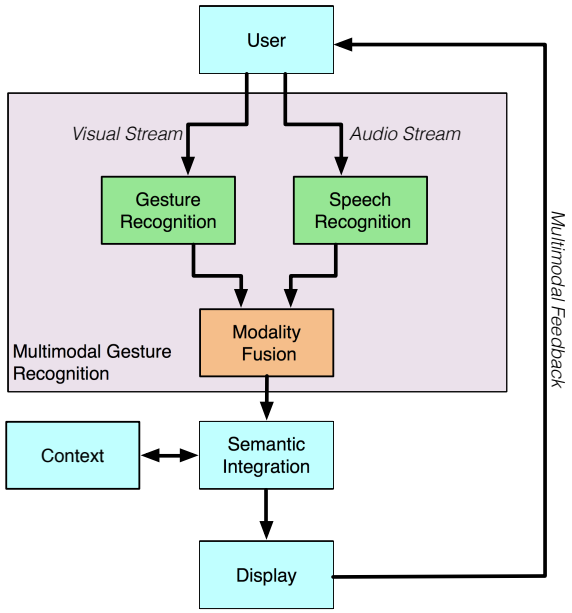


Figure 11.5 **Multimodal gesture recognition integrated into an HCI system.** (Based on a figure by Krahnstoever et al. [2002])

Temporal modeling is imperative, among others, for disambiguating patterns of motion, which otherwise have very similar properties, e.g., a "come-here" gesture vs. a "go-away" one, or "sit-down" vs. "stand-up". Over the years, various temporal models have been proposed for this purpose, with hidden Markov models (HMMs) and their variants being the prevalent ones until recently [Chen et al. 2003, Mitra and Acharya 2007, Pisharady and Saerbeck 2015, Wilson and Bobick 1999]. Gesture HMMs statistically model sequences of feature vectors (observations) that are extracted from gesture data, such as static geometric or spatial features (e.g., shape descriptors) or dynamic ones (e.g., optical flow). Their effectiveness is based on the assumption that observations at any instant can be modeled as dependent on the current hidden state only. To train HMMs, the expectation-maximization algorithm is applied, iteratively updating model parameters to maximize the observed data likelihood [Rabiner 1989]. This makes HMMs generative rather than discriminative in nature, which in turn implies that they are not optimal for classification or recognition tasks.

Being more appropriate in that respect, discriminative models based on hidden conditional random fields (HCRFs) [Morency et al. 2007, Wang et al. 2006] were shown to outperform HMMs in gesture recognition. Support vector machines (SVMs), applied on a sliding window of the observations, have also demonstrated competitive performance using complex feature extraction schemes, such as dense trajectories [Peng et al. 2015], bag-of-words representa-

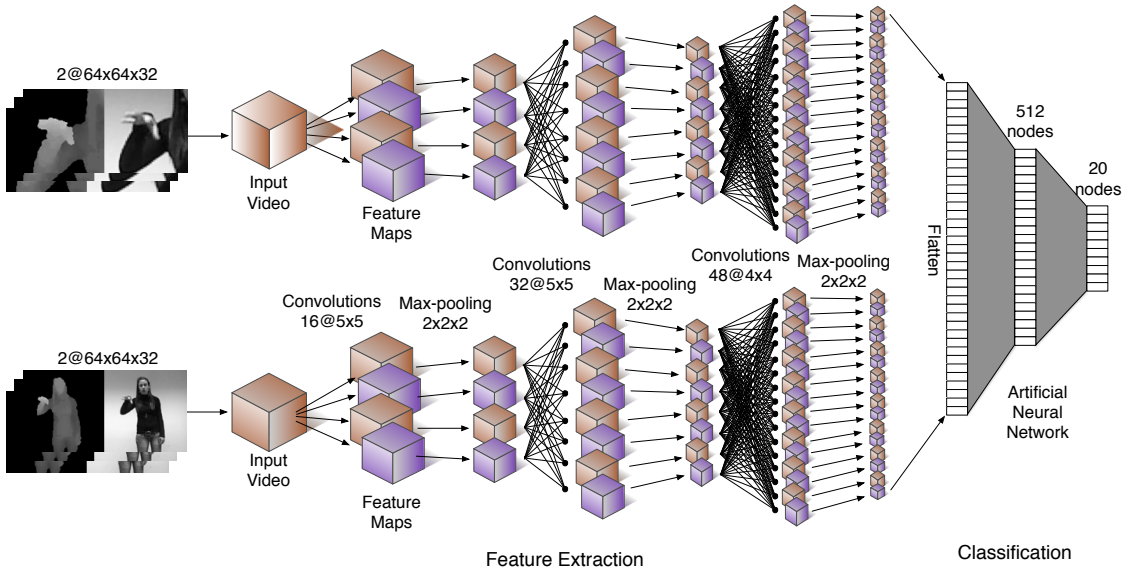


Figure 11.6 **Architecture of the CNN-based gesture recognition system of Pigou et al. [2015].** Two CNNs are used in parallel to extract features from RGB-D data of the primary hand (upper part of the diagram) and body (lower part) over 32-frame long data sequences. Convolutional and max pooling layers alternate at the feature extraction stage, and gesture classification is performed by a three-layer, feed-forward, fully-connected ANN. The max-pooling region, as well as the number and size of CNN filters are also shown, e.g., 16 filters of size 5x5 are employed at the first convolutional layer. (Based on a figure by Pigou et al. [2015])

tions [Laptev et al. 2008], and poselets [Bourdev and Malik 2009]. However, HMMs, SVMs, and HCRFs, being essentially single-layer models, are not able to learn higher-level representations and to fully benefit from large corpora of gestures becoming increasingly available.

To address this shortcoming, the use of deep learning is becoming increasingly popular in a number of spatio-temporal video-based classification tasks, including gesture and action recognition. For example, Taylor et al. [2010] extract features from traditional RGB videos using convolutional neural networks (CNNs), instead of hand-crafting the features. Similarly, Pigou et al. [2015] employ ***RGB-D*** data and apply two separate CNNs to extract features for the entire body and for the hands. These are then fed to an artificial neural network (ANN) for gesture classification, as depicted in Figure 11.6. Three-dimensional CNNs (3D-CNNs) are introduced by Ji et al. [2013] for recognizing human actions in surveillance videos. These models essentially perform 3D convolutions along both spatial and temporal dimensions, by

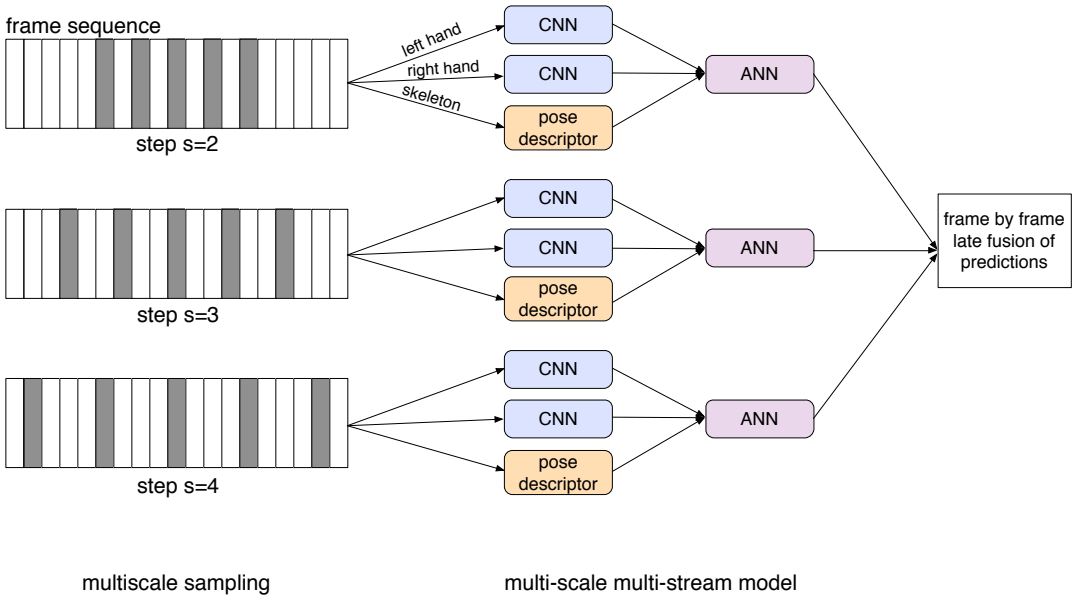


Figure 11.7 **Architecture of the multi-scale, multi-stream gesture recognition system of Neverova et al. [2015].** Input frame sequences are sampled at three temporal scales, i.e., every 2, 3, or 4 frames (top-to-bottom). At each scale, RGB-D image sequences of the left and right hands are fed into separate CNNs, and their outputs are combined using a feed-forward ANN (similarly to Figure 11.6). Skeleton-based features are also fed into the ANNs that generate scale-dependent gesture recognition results. These are subsequently combined by a weighted averaging scheme to yield the final result. (Based on a figure by Neverova et al. [2015])

convolving the cube formed by adjacent 2D frames with a 3D kernel. This way, 3D-CNNs capture motion information in a sequence of contiguous frames. In [Wu et al. 2016], a 3D-CNN is applied on RGB-D data, and a deep belief network is used to model dynamics of coordinates of human skeleton joints. The two networks are fused at the output stage by means of a linear combination of their output posterior probabilities. These are then used as observation probabilities of gesture HMMs to model higher-level temporal relationships, allowing continuous gesture segmentation and recognition.

Stream-dependent CNNs are also considered by Neverova et al. [2015], this time at multiple temporal scales, as depicted in Figure 11.7. Instead of using a temporal model, this approach is window-based, i.e., classification is performed employing a sliding window over

the sequence of observations. Fusion across temporal scales occurs at the top level of the proposed architecture by means of weighted averaging to yield the final result, whereas multi-stream fusion is performed at a deeper level.

Data Resources and Challenges for Gesture Recognition

Concerning data resources for gesture recognition, the recent publication of the ChaLearn databases, collected using RGB-D or ***RGB-D-A*** sensors, and the corresponding Challenges [Escalera et al. 2016] have provided a major evaluation benchmark for state-of-the-art gesture recognition approaches (examples from one of these sets are shown in Figure 11.3).

In particular, the first ChaLearn Challenge and corpus concerned multimodal gesture recognition [Escalera et al. 2013b]. Most of the developed systems on these data were based on HMMs [Pitsikalis et al. 2015], but other approaches such as neural networks, SVMs, and random forests were also applied. For the visual recognition of gestures, low-level features such as space-time interest points [Willems et al. 2008] were used by Nandakumar et al. [2013], and, to model large-scale temporal dependencies, recurrent neural networks (RNNs) were applied by Neverova et al. [2013]. The best performing system in the Challenge, proposed by Wu et al. [2013], employed dynamic time warping (DTW) [Rabiner and Juang 1993] to achieve visual gesture classification based on features extracted from the human body skeleton.

The next ChaLearn Challenge (Looking at People – LAP) and gesture recognition corpus [Escalera et al. 2015] targeted gesturer-independent classification of gestures. The best approaches in the corresponding task were based on sophisticated hand-crafted visual features for the available input streams. For example, four types of skeleton features were extracted at multiple temporal scales by Monnier et al. [2015], namely normalized positions of the upper body major joints, joint quaternion angles, Euclidean distances between specific joints, and directed distances between pairs of joints, following Yao et al. [2011]. Derivatives of all the above were used along with handshape features based on histograms of oriented gradients (HOGs), originally proposed by Dalal and Triggs [2005]. The first ranked team [Neverova et al. 2015] employed a descriptor comprising seven subsets of skeleton features. More specifically, the adopted representation included joint velocities, positions, accelerations, as well as inclination and azimuth angles.

11.3.2 Speech Recognition

Deep learning has recently prevailed in the field of automatic speech recognition (ASR) as well. The remarkable parallel computing capabilities of modern graphics processing units, the introduction of effective training algorithms, and the abundance of available speech data have allowed deep architectures to unfold their potential. Indeed, early works demonstrated their superior performance in both phone recognition [Mohamed et al. 2009] and large-vocabulary continuous speech recognition [Dahl et al. 2012], as compared to HMMs with

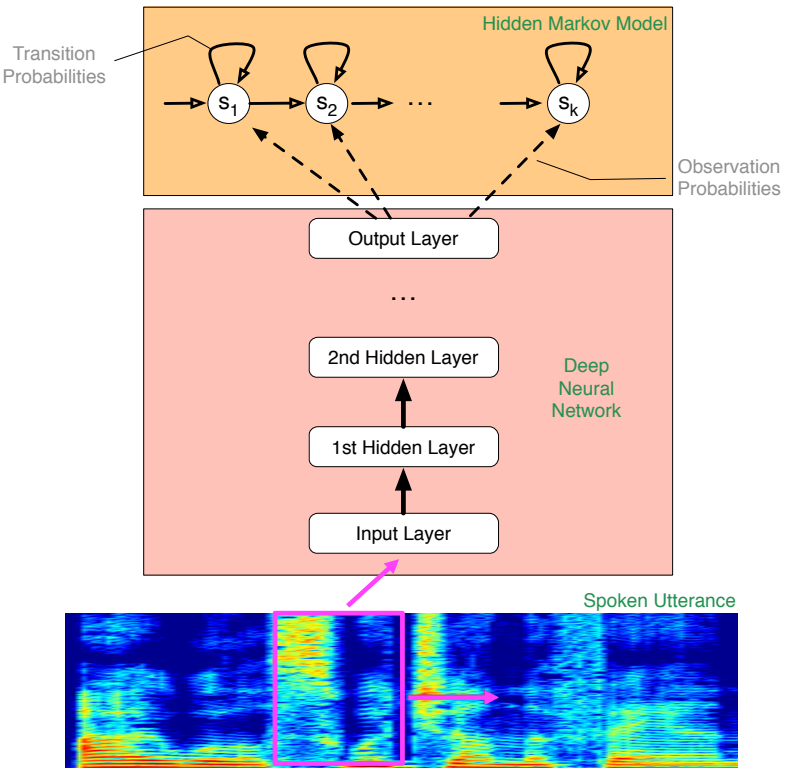


Figure 11.8 **Typical DNN-HMM architecture for speech recognition.** Filtered spectral energies are used at the input DNN layer. Observation probabilities at each HMM state are estimated using the DNN output posteriors. HMM states typically correspond to context-dependent sub-phonemes [Yu and Deng 2011].

conventional observation probabilities based on Gaussian mixture models (GMMs), known as GMM-HMMs. Since then, deep neural networks (DNNs) and their variants have become indispensable in state-of-the-art ASR systems.

In fact, the idea to combine ANNs and HMMs to improve ASR first appeared almost 30 years ago. More specifically, the most promising ANN-HMM approaches back then would either use ANNs to estimate posterior probabilities for each HMM state ("connectionist" approach) [Bourlard and Morgan 1994, Morgan and Bourlard 1990], or process these posterior probabilities appropriately and treat them as features in a GMM-HMM system ("tandem" approach) [Hermansky et al. 2000]. In both cases, the results would not necessarily outperform the rival systems based on conventional GMM-HMMs that were being very intensely developed in parallel. Despite the very appealing properties of the ANN-HMM approach, its

inherently discriminative nature and modeling effectiveness among others, it was not possible to effectively train systems with more than two hidden ANN layers. Insufficient computational resources, the lack of large enough corpora, and limitations of the applied training algorithms were the main culprits. However, all these obstacles have recently been overcome, ANNs have become DNNs, and DNN-HMMs consistently demonstrate the best performance in all major ASR tasks [Hinton et al. 2012].

A characteristic architecture of a DNN-HMM speech recognition system is depicted in Figure 11.8. Speech features typically comprise filtered spectral energies at various frequency bands, and the number of DNN hidden layers may vary, depending on available data and task specifics [Yu and Deng 2011]. Alternative types of networks and combinations, each accepting either the same or appropriately transformed speech input, have also been successfully proposed, e.g., CNNs, RNNs, and long short-term memory networks, leading to state-of-the-art performance in a number of challenging ASR tasks [Hannun et al. 2014, Saon et al. 2016, Xiong et al. 2017]. Research in this area is currently extremely active, and further improvements are expected. Among others, such could concern training algorithms, multi-task learning, i.e., re-using models trained for a particular task in a different one, and handling cases where limited only data are available [Deng et al. 2013].

11.3.3 Multimodal Fusion

Correct interpretation of user intent based on speech and accompanying gestures requires appropriate fusion of the two modalities. Fusion, or multimodal integration, is probably the most critical component of a multimodal gesture recognition system. In natural interactions, even if the results of the single-modality recognition modules may be quite reliable, their integrated interpretation can be difficult to determine. For example, certain co-occurrences can be interpreted as redundant expressions of the same meaning, like “bye bye” and the corresponding emblematic gesture, while others need to be combined, e.g., “go there” and the accompanying pointing gesture (see also the distinction between redundant and complementary speech and gestures in [Kopp and Bergmann 2017]).

Related to this, an illustrative classification of multimodal interfaces is provided in Table 11.1, based on the corresponding multimodal interpretation and on the relative timing of event occurrences in the involved modalities [Nigay and Coutaz 1993]. In the synergistic case, both speech and gestures are available in parallel, and they must be fully integrated to actually interpret user intent. It should be pointed out that “parallel” does not necessarily mean simultaneous; empirical studies have shown that expecting the user’s speech and corresponding gestures to fully overlap is an oversimplification [Oviatt and Cohen 2015]. Given this finding, a synergistic interface would expect the “put that there” command to be delivered by the user synchronizing the first pointing gesture with the utterance of “put that” and the second pointing gesture with the utterance of “there”. The exact expected timing between the delivery of

speech and gestures should ideally be adapted to the user and the context (see [Huang et al. 2006] for a machine learning approach to that).

In the concurrent scenario, speech and gestures are independent, but occur in parallel. The user, for example, can increase the brightness of the monitor using speech, while, at the same time, zooming-in with a pinching gesture. On the other hand, concerning interfaces that expect sequential utilization of the two modalities, the alternate case refers to when input is provided in one of the modalities and is then complemented via the other, while the exclusive case to when the events in the two modalities are handled separately. An alternate multimodal interface would require, for example, saying "put that there" first, and then sequentially performing two pointing gestures (one for "that" and one for "there"). The synergistic and the alternate modes of operation are the ones we are mostly interested in, but there are potential benefits in using the other types as well [Turk 2014].

For synergistic systems, the next question that needs to be addressed regards the level at which fusion should be performed. To answer this, a comprehensive recent survey of multimodal integration approaches [Lalanne et al. 2009] uses the seven-layered protocol model of HCI [Nielsen 1986]. This model accounts for the physical, perceptual, and conceptual representations involved, with multimodal fusion potentially feasible at any of its seven layers. An example is provided in Table 11.2, for the case of a user saying "remove this" and pointing to a specific shape on the screen, with the goal of removing a table from a virtual room. The corresponding representations at each layer of the interaction are shown.

More specifically, fusion of multiple modality feature streams at the physical world layers of Table 11.2 is often referred to as "early integration". Such approach typically requires that the involved modalities are tightly synchronized, as for example in the case of audio-visual ASR [Potamianos et al. 2017]. Since this is not guaranteed in speech- and gesture-based interaction, early integration is not directly applicable to multimodal gesture recognition, at least following the paradigms that have appeared in the literature so far. Instead, fusion is typically applied at the perceptual or the conceptual world layers of Table 11.2, being often referred to as "intermediate" and "late integration", respectively [Maragos et al. 2008]. Such approaches are also commonly employed in pen and speech interfaces, as discussed in Chapter

Table 11.1 Multimodal interface types based on the corresponding multimodal interpretation and on the relative timing of the events occurring in each modality.

		Modality Occurrence Timing	
		Sequential	*Parallel*
Multimodal Interpretation	*Combined*	Alternate	Synergistic
	Independent	Exclusive	Concurrent

Table 11.2 **The seven-layer model of HCI [Nielsen 1986].** The model has been adapted to multimodal gesture recognition, in the example of a user saying "remove this" and pointing to a specific shape on the screen, with the goal of removing a table from a virtual room. Multimodal fusion can potentially occur at any of the seven layers depicted. Further details are discussed in Section 11.3.3.

Level	Type	Units	Definition	Example	World
1	Goal	Real-world concepts	Mentalization of a goal, a wish in the user's head	Remove the table from the virtual room on the screen	Conceptual
2	Pragmatic	System concepts	Translation of a goal into system concepts	Delete the table that lies at the particular point of the room	
3	Semantic	Detailed functions	Real world objects translated into system objects manipulated by functions	Delete object in screen coordinates	
4	Syntactic	System sentences	Time and space sequencing of information units	Delete screen coordinates	Perceptual
5	Lexical	Information units	Smallest elements transporting significant information: word, figure, screen coordinates, icon	[delete] command, screen coordinates	
6	Alphabetic	Lexemes	Primitive symbols: letters, numbers, columns, lines, dots, phonemes, ...	/r/, /iy/, /m/, /uw/, /v/, right-hand pointing index finger	Physical
7	Physical	Physically coded information	Light, sound, physical movement	Saying "remove this", pointing to a shape on the screen	

10 of this volume [Cohen and Oviatt 2017]. For example, Johnston and Bangalore [2000] use a multimodal weighted finite-state transducer to generate the final decision based on speech and pen pointing recognition. In other work, Johnston [1998] applies a unification-based method for fusion, where speech and pen drawings are represented by attribute-value structures and integrated into single multimodal representations by means of appropriate grammars. Further, Wu et al. [1999] propose a statistical fusion approach, where posterior probabilities of the speech and pen drawing or pointing recognition results are appropriately integrated, and the multimodal combination with the highest probability is selected as the final output. Such fusion approaches are directly transferable to multimodal gesture recognition as well.

More elaborate statistical approaches for fusion at the lexical layer for multimodal gesture recognition are presented in [Miki et al. 2014, Pitsikalis et al. 2015]. There, N-best lists

of symbol sequences are generated using the two modalities separately, and then properly rescored based on both. The best-scoring hypothesis is subsequently returned as the final multimodal gesture recognition result. Additional details on this approach are provided in Section 11.4.2. Further, a framework to more tightly integrate the available modalities, even during the training phase, is proposed by Wu and Cheng [2014]. Building on boosting learning [Zhang and Yu 2005] and co-training [Blum and Mitchell 1998], this approach iteratively combines at the lexical level many co-trained weak classifiers for all modalities to construct a multimodal strong classifier. The method is shown to achieve state-of-the-art performance in a number of multimodal gesture recognition tasks.

11.4 A System in Detail

In this section, we highlight the main ideas and methods of the authors' recent work on multimodal gesture recognition [Pitsikalis et al. 2015]. The developed system comprises two main modules: feature extraction and modeling for each of three available unimodal data streams, as overviewed in Section 11.4.1, and multimodal fusion that is discussed in Section 11.4.2. The presentation aims to facilitate a better understanding of the practical aspects of addressing these tasks.

System development and evaluation utilizes the demanding ChaLearn dataset [Escalera et al. 2013b] that focuses on multiple-instance, user-independent learning of gestures from multimodal data. The corpus was used in the multimodal gesture recognition Challenge, held in conjunction with the ICMI'13 conference [Escalera et al. 2013a], and is available online [ChaLearn 2013] (samples of it are depicted in Figure 11.3). The data are captured using an RGB-D-A sensor (Kinect) providing a number of data streams, namely sequences of both RGB and depth image frames of the gesturer's body, skeleton information, orientation of body joints, as well as concurrently recorded audio of the accompanying speech, essentially having the same meaning as the performed gesture.

Each recording in the ChaLearn corpus may include a sequence of gestures (ten or more), and the goal is to recognize the entire sequence as accurately as possible. The vocabulary used contains twenty Italian cultural-anthropological gestures. Examples include, among others, utterances "come here", "I am hungry", and "go away" in Italian, accompanied by their corresponding gestural expressions. The dataset is partitioned into three subsets, used for development, validation, and final evaluation, including 39 users and approximately 14k multimodal gesture instances in total, all manually transcribed and loosely end-pointed. The corresponding temporal boundaries are provided as part of the corpus, and they are exploited in the training phase of the presented system.

There are several issues that render multimodal gesture recognition on this dataset quite challenging, as described by [Escalera et al. 2013b]. These include the requirement to process long continuous sequences of gestures (i.e., not just isolated instances), as well as the need to account for a relatively large vocabulary of gestures and a large variety of users. Further,

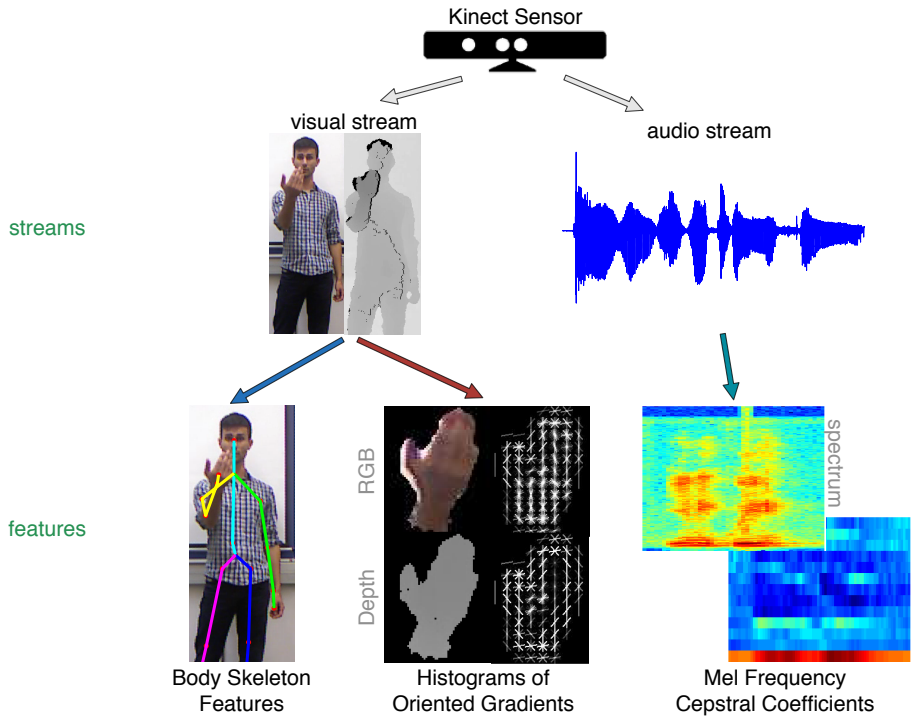


Figure 11.9 **Schematic depicting the feature extraction pipeline of the multimodal gesture system, detailed in Section 11.4.1.** Top row: Kinect sensor providing the data streams; Middle row: visual data (RGB and depth) and audio data; Bottom row: visual features (skeleton-based, as well as HOGs extracted from the RGB and depth channels) and audio features (MFCCs).

the presence of distracting, non-sense (non-target) gestures and spoken phrases, referred to as out-of-vocabulary (OOV) for both modalities, makes the task even harder.

11.4.1 Single-Stream Feature Extraction and Modeling

The developed multimodal gesture recognition system exploits both raw audio-visual and human skeleton metadata streams provided by the Kinect. Such data are processed to yield three feature streams, namely: one extracted from the skeleton information (visual metadata); a second containing visual features from both depth and planar RGB video; and a third exploiting the audio data. The extracted features are briefly described in the following and are schematically depicted in Figure 11.9:

- *Skeleton:* Skeletal features are based on the 3D coordinates of the body skeletal points, readily provided by the Kinect sensor [Shotton et al. 2013]. More specifically, the employed feature set includes 27 features in total: the 3D positions of hands and elbows, the relative 3D positions of hands with respect to the corresponding elbows, the 3D direction of hand movement, and the 3D distance between the hand centroids.
- *Handshape:* The handshape is represented by means of HOG features, which are visual descriptors popular in the computer vision literature [Buehler et al. 2009]. First, segmented images of both hands in the planar RGB video and depth data streams are obtained, exploiting the hand tracking results provided by the Kinect and using simple depth-based thresholding. Subsequently, HOGs are extracted on both RGB and depth hand images, employing a 7×7 grid and 4-bin histograms, using the open-source VLFeat library [Vedaldi and Fulkerson 2008]. The resulting feature vector dimensionality is 392. Both handshape and skeleton features are extracted at 20 frames per second.
- *Audio:* To efficiently capture the spectral properties of speech signals, the acoustic frontend generates 39-dimensional feature vectors every 10 msec. Each such vector comprises 13 Mel Frequency Cepstral Coefficients (MFCCs), along with their first- and second-order temporal derivatives.

Following feature extraction, the underlying single-stream modeling scheme is based on HMMs, building on the keyword-filler paradigm that was originally introduced for speech [Rose and Paul 1990, Wilpon et al. 1990] in applications like spoken document indexing and retrieval [Foote 1999] and speech surveillance [Rose 1992]. Similarly to these works, recognition of a specific set of gestures within an observed sequence, comprising other heterogeneous events as well, is viewed as a keyword detection problem. The twenty gestures to be recognized are the keywords, and all the rest is ignored. Then, for every information stream, each gesture, or in practice its projection on that stream (considering gestures as multimodal events), is modeled by an HMM. In addition, for each stream, a separate HMM is trained to represent silence/inactivity, and another one, referred to as the background model (BM), accounts for all OOVs appearing in that stream. All these models are three-state, left-to-right HMMs with GMMs employed as state-dependent observation probability distributions. The actual number of mixture components used in the experiments is optimized on the development set. Given enough training data, DNNs could also be used for this purpose. The proposed modeling scheme is implemented using the open-source hidden Markov model toolkit [HTK 2009, Young et al. 2002].

11.4.2 Multimodal Fusion

Gestures in the ChaLearn dataset occur in parallel to their semantically corresponding verbal expressions. By fusing information from all available modalities, one can improve robustness of the targeted multimodal interface. As a side note, based on the classification scheme pre-

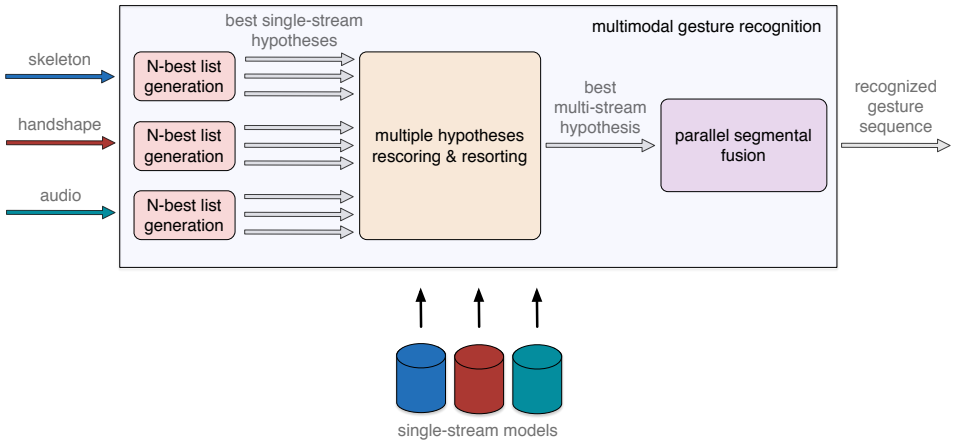


Figure 11.10 **Overview of the multimodal fusion scheme for gesture recognition, detailed in Section 11.4.2.** Single-stream models are first used to generate possible hypotheses for the observed gesture sequence. These are subsequently rescored by all streams, and the best one is selected. Finally, a parallel segmental fusion scheme is applied to yield the result.

sented in Table 11.1, such an interface would be synergistic: although the available modalities convey the same meaning and do not complement each other, they synergize to improve recognition robustness. Given a vocabulary of multimodal gestures and a sequence of concurrent observations from multiple input streams, the goal is to obtain the best possible hypothesis regarding the observed gesture sequence. In the presented system, the final decision is based on the three feature streams discussed in Section 11.4.1, namely skeleton-based, handshape features, and auditory spectral ones. In essence, any set of information streams can be employed within this framework.

More specifically, a two-level fusion approach is implemented to combine information from the individual feature streams:

- *1st Pass (P1):* First, the single-stream models described in Section 11.4.1 are used to generate a set of possible gesture sequence hypotheses for the given observations. This initial set of hypotheses is then multimodally rescored and re-sorted.
- *2nd Pass (P2):* Based on the temporal boundaries of the gestures in the best hypothesis after P1, the parallel segmental fusion scheme of Vogler and Metaxas [2001] is applied to further improve recognition performance.

The approach is schematically depicted in Figure 11.10. A more detailed description of the fusion steps is provided next.

N-Best Rescoring and Re-sorting

Using the single-stream gesture models and a grammar that defines the set of alternative sequences of gestures allowed (essentially any gesture can follow any other, with silence and OOVs possibly appearing in-between), a list of N-best hypotheses is initially generated for the unknown sequence for each stream. To achieve this, a variation of the Viterbi algorithm is applied, namely the lattice N-best algorithm [Schwartz and Austin 1991], which apart from storing just the single best gesture at each node, it also records additional highly-scoring gestures together with their scores.

The superset of all these N-best hypotheses, after removing duplicates, is subsequently rescored and re-sorted. The combined new score for each hypothesis is a weighted sum of single-stream scores after normalization, i.e., after transforming them to have zero mean and standard deviation equal to one (across each stream hypothesis scores). Stream weights are determined experimentally on a held-out validation set.

More specifically, the stream-based scores are estimated by means of "force-aligning" each gesture sequence hypothesis with the observation sequence in the particular stream. This is achieved by means of the Viterbi algorithm [Rabiner and Juang 1993] applied in a constrained search space, so that it can only output alternative temporal alignments for the specific hypothesis. Insertions or deletions of silence and possibly other OOVs are also accounted for, in order to allow, for example, a hypothesis originally based on audio to better match the visual observation sequence, e.g., when the user randomly moves the hands without saying anything. The hypothesis with the maximum combined score is returned and passed to the second fusion step, detailed next.

Segmental Parallel Fusion

At this step, the temporal boundaries of the best hypothesis mentioned above are used to segment the observation streams and reduce the recognition problem to a segmental classification one. It has to be mentioned that temporal boundaries may be different for each stream, since temporal synchronization is not guaranteed. For every segment, corresponding to a detected gesture, and for each stream, the log-likelihood of all stream-based gesture models is estimated given the corresponding observation sequence. The segmental scores are then linearly combined across modalities to get a multimodal score of all possible gestures. The parameters of this linear combination are determined experimentally on a held-out validation set. Finally, for each segment, the recognized gesture is the one with the highest multimodal score. This final step is expected to yield additional improvements and correct false alarms by seeking loosely overlapping multimodal evidence in support of each hypothesized gesture.

11.4.3 Experimental Results

As already mentioned, the performance of the proposed multimodal gesture scheme has been evaluated on the ChaLearn corpus [Pavlakos et al. 2014, Pitsikalis et al. 2015]. In Table 11.3, recognition results are presented for each stream separately, as well as after single- or two-stage fusion. All results are expressed in accuracy (%), computed as 100 – WER, where WER is the percent word error rate that accounts for insertions, deletions, and substitutions. Based on the audio modality alone, recognition accuracy reaches 87.2%, while the second best-performing modality is the skeleton (49.1% accuracy). Combining all modalities by means of segmental parallel fusion (P2) alone improves performance to 88.5%, while multimodal N-best rescoring (P1) alone leads to 93.1% accuracy (in the presented experiments, N was chosen equal to 200). To evaluate the P2 component separately, no multimodal rescoring of hypotheses was performed, and the segments in all modalities were determined after forced-alignment to the best audio-based hypothesis. The proposed multimodal fusion scheme employing both P1 and P2 yields the best recognition accuracy (93.3%), reducing error by a relative 48% over a baseline approach discussed next.

For comparison purposes in Table 11.3, the best performing approach at the ChaLearn multimodal gesture recognition Challenge [Escalera et al. 2013b, Wu et al. 2013] is considered as the baseline. That system only employs the audio and skeletal input streams. First, a simple time-domain end-point detection algorithm based on coordinates of body joints is applied to split continuous data sequences into candidate gesture segments. Then audio HMMs (similar to the models of the presented system) are used to estimate an audio-based score for each gesture in each segment. In an analogous manner, a DTW-based skeletal feature classifier is applied to estimate corresponding visual-based scores. These single-stream scores by the two classifiers are normalized and then combined to produce a weighted sum for fusion. All in all, the baseline system resembles the use of the P2 fusion step alone, also performing quite similarly to it. Given that, the presented system leads to better performance, mainly because it also accounts for the N-best hypotheses and not just the best one (via integration of the P1 fusion step).

Table 11.3 Evaluation results on the ChaLearn corpus of the various components of the multimodal gesture recognition system of Section 11.4.

Unimodal	Accuracy (%)	Multimodal	Accuracy (%)
Handshape	20.2	Fusion – P2	88.5
Skeleton	49.1	Fusion – P1	93.1
Audio	87.2	Fusion – P1+P2	**93.3**
		Baseline	87.2

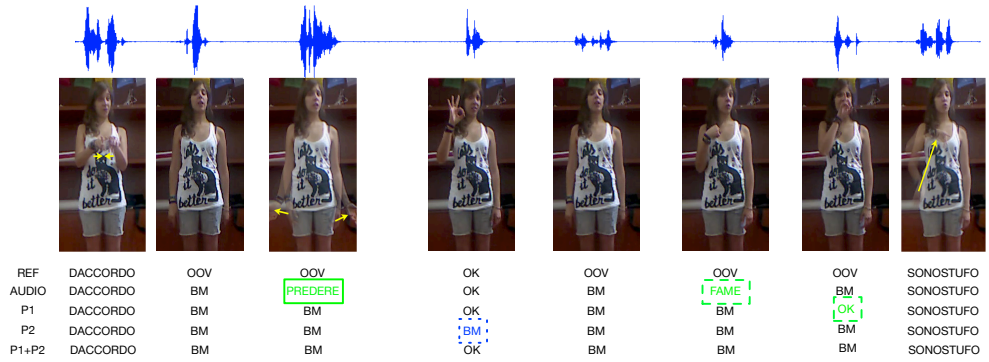


Figure 11.11 **Example recognition results on a ChaLearn corpus sequence, employing the techniques of Section 11.4.** Input audio and visual streams are depicted in the first and second rows, respectively. Ground-truth annotations are given in the row labeled as "REF". Speech recognition results are shown in row "AUDIO". Finally, multimodal gesture recognition results using fusion schemes P1, P2, and P1+P2 are depicted in the last three rows. Deletion errors are highlighted in blue and insertions in green. A corresponding video demonstration can be found at: https://www.youtube.com/watch?v=xrsRsxz3Vk4 (From [Pavlakos et al. 2014])

A gesture sequence recognition example is shown in Figure 11.11. Both audio and visual modalities are illustrated for a gesture sequence, accompanied by the ground-truth gesture-level transcriptions (row: "REF"). In addition, the recognition output employing the audio modality alone (AUDIO) is shown, along with the three presented fusion schemes (P1, P2, and P1 + P2). Focusing on audio-only recognition results, one notices two insertions (words "PREDERE" and "FAME"). By employing either the P1 or P2 schemes separately, the above gesture insertions are corrected, as the use of the visual modality helps identifying that these segments correspond to OOVs. However, recognition is still not perfect, since other insertion or deletion errors arise. On the contrary, the application of the proposed two-step fusion scheme leads to the best possible recognition result in this case.

In general, correct identification of OOVs and elimination of false alarms remains challenging. To address this problem, it would be beneficial to explicitly add modality synchronization constraints and, in this way, properly detect events that are not really multimodal, i.e., the majority of OOVs in ChaLearn. Further, as discussed earlier, properly adapting the proposed methodology to employ discriminative single-stream models, e.g., DNN-HMMs, instead of GMM-HMMs, can lead to additional performance improvements.

11.5 Conclusions and Outlook

In this chapter, after introducing the basic properties of gestures and discussing the various forms of speech-gesture interplay, we presented an overview of the current trends in the field of multimodal gesture recognition. We specifically focused on recent advances in gesture recognition, speech recognition, and multimodal fusion, also providing up-to-date references for the interested reader to delve deeper into the presented methodologies. Last, we described in detail a state-of-the-art multimodal gesture recognition system that employs a two-stage fusion scheme, evaluating it on a well-known multimodal gesture dataset.

Multimodal gesture recognition has definitely come a long way since the seminal work of Bolt [1980]. The increasing availability of RGB-D-A sensors like the Kinect has allowed the widespread collection of gesture and speech data and has significantly boosted related research. The application of the deep learning paradigm has already led to significant improvements in the speech and gesture recognition domains separately, and recently proposed fusion approaches also bear significant potential to push developed systems a little closer to real-life applications. Multimodal gesture recognition systems are already foreseen in ambient assisted living [Rodomagoulakis et al. 2016], gaming applications [Liu and Kavakli 2010], and learning environments [Miki et al. 2014], among others. In this direction, however, there are numerous challenges remaining to be addressed. Automatic gesture and speech recognition cannot yet be considered as solved problems, especially in spontaneous natural interaction. Similarly, fusion approaches cannot yet easily overcome the difficulties emerging in real-life scenarios, where errors, delays, and large variability are commonplace. Towards overcoming these hurdles, the ongoing collection of annotated, representative, "in the wild" corpora of speech and gestures, the development of standardized benchmarks, as well as the increasingly broader availability and improvement of open-source software toolkits for recognition bear the promise to further promote research in the area and lead to significant advances in the near future.

Acknowledgments

Writing of this chapter was supported by the Onassis Foundation and by the European Union under research projects BabyRobot (grant H2020-687831) and I-SUPPORT (grant H2020-643666). The authors are grateful to the editors and reviewers for their valuable comments and suggestions. They also wish to thank Georgios Pavlakos at University of Pennsylvania for his contributions to an early version of the proposed multimodal gesture recognition system.

Focus Questions

11.1. Record a video of a face-to-face discussion you have with a friend. After you carefully examine the recording, identify the co-speech gestures you have performed. Which are the roles played by the gestures you have identified? Are they just communicative, or do

they also facilitate more efficient speech production? Then, do the same for a video of yourself while having a telephone conversation. How does your gesturing differ in the two cases?

11.2. Have you tried to participate in a face-to-face discussion and consciously avoid gesturing? How do you think your speech was or would be affected?

11.3. Using the video recordings described above, try to categorize all identified co-speech gestures using the scheme detailed in Section 11.2.1. Are there gestures that would possibly fit in two or more categories?

11.4. Based on your experience of human-computer interfaces, can you identify a number of gestures and/or verbal expressions that you typically use when interacting with an electronic device? Have you experienced multimodal interaction with a computer? Can you think of a specific use-case in which a multimodal human-computer interface would really be beneficial? Based on the classification scheme in Table 11.1, what type of interface is that?

11.5. How are the temporal dynamics of gestures typically modeled in a gesture recognition setup? How are dynamics typically treated differently by HMMs in comparison to SVMs or ANNs?

11.6. To what extent are state-of-the-art approaches for speech and gesture recognition similar? In which aspects do they differ?

11.7. Based on Table 11.2, and similarly to the example already provided there, can you think of another example of multimodal human-computer interaction and identify how information proceeds through all seven layers, i.e., from the physical one all the way up to the ultimate goal level? Where do you think fusion would be more appropriate in your example?

11.8. How would the proposed multimodal gesture recognition approach of Section 11.4 differ if DNNs were employed instead? How can DNNs benefit the proposed multimodal gesture recognition scheme?

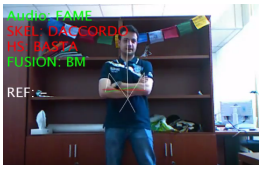
Chapter Digital Resources (Videos)

Video thumbnails:

Video 11.1
(Fig.11.1)

Video 11.2
(Fig. 11.4a)

Video 11.3
(Fig. 11.4b)



Video 11.4
(Fig. 11.11
thumbnail)

Video 11.1. Digital resource for Figure 11.1.
Video of the "Put that there!" prototype by the Architecture Machine Group at MIT (slighly shortened version of the original,
available at: https://www.youtube.com/watch?v=RyBEUyEtxQo).

Video 11.2. Digital resource for Figure 11.4a.
A demonstration video of a gesture interface to control computer functionality.
Available at: https://www.youtube.com/watch?v=Wq1FM84uAck

Video 11.3. Digital resource for Figure 11.4b.
A demonstration video of gesture recognition in a virtual reality environment.
Available at: https://www.youtube.com/watch?v=ALBsLEupolY

Video 11.4. Digital resource for Figure 11.11.
A demonstration video of the authors' multimodal gesture recognition system presented in Section 11.4. Available at: https://www.youtube.com/watch?v=xrsRsxz3Vk4

Bibliography

A. Blum and T. Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the Annual Conference on Computational Learning Theory (COLT)*, pp. 92–100.

R. A. Bolt. 1980. Put-that-there: Voice and gesture at the graphics interface. *ACM SIGGRAPH Computer Graphics*, 14(3): 262–270.

L. Bourdev and J. Malik. 2009. Poselets: Body part detectors trained using 3D human pose annotations. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1365–1372.

H. A. Bourlard and N. Morgan. 1994. *Connectionist Speech Recognition: A Hybrid Approach*, volume SECS 247. Springer Science & Business Media, New York, NY.

J. Bressem and S. H. Ladewig. 2011. Rethinking gesture phases: Articulatory features of gestural movement? *Semiotica*, 2011(184): 53–91.

P. Buehler, A. Zisserman, and M. Everingham. 2009. Learning sign language by watching TV (using weakly aligned subtitles). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2961–2968.

J. Cassell. 1998. A framework for gesture generation and interpretation. In R. Cipolla and A. Pentland, eds., *Computer Vision in Human-Machine Interaction*, pp. 191–215. Cambridge University Press, New York, NY.

J. Cassell, D. McNeill, and K.-E. McCullough. 1999. Speech-gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information. *Pragmatics & Cognition*, 7(1): 1–34.

ChaLearn, 2013. http://chalearnlap.cvc.uab.es/dataset/12/description/. (last accessed January 6, 2017).

A. Chaudhary, J. L. Raheja, K. Das, and S. Raheja. 2011. Intelligent approaches to interact with machines using hand gesture recognition in natural way: A survey. *International Journal of Computer Science & Engineering Survey*, 2(1): 122–133.

P. Chawla and R. M. Krauss. 1994. Gesture and speech in spontaneous and rehearsed narratives. *Journal of Experimental Social Psychology*, 30(6): 580–601.

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

P. R. Cohen and S. Oviatt. 2017. Multimodal speech and pen interfaces. In S. Oviatt, B. Schuller, P. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, eds., *The Handbook of Multimodal-Multisensor Interfaces, Volume 1: Foundations, User Modeling, and Multimodal Combinations*. Morgan Claypool Publishers, San Rafael, CA.

J.-M. Colletta, C. Pellenq, and M. Guidetti. 2010. Age-related changes in co-speech gesture and narrative: Evidence from French children and adults. *Speech Communication*, 52(6): 565–576.

G. E. Dahl, D. Yu, L. Deng, and A. Acero. 2012. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language*

Processing, 20(1): 30–42.

N. Dalal and B. Triggs. 2005. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pp. 886–893.

J. P. de Ruiter. 2006. Can gesticulation help aphasic people speak, or rather, communicate? *Advances in Speech Language Pathology*, 8(2): 124–127.

L. Deng and D. Yu. 2014. Deep learning: Methods and applications. *Foundations and Trends® in Signal Processing*, 7(3–4): 197–387.

L. Deng, G. Hinton, and B. Kingsbury. 2013. New types of deep neural network learning for speech recognition and related applications: an overview. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8599–8603.

D. Efron. 1941. *Gesture and Environment*. King's Crown Press, New York, NY.

P. Ekman and W. V. Friesen. 1969. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica*, 1(1): 49–98.

S. Escalera, J. Gonzàlez, X. Baró, M. Reyes, I. Guyon, V. Athitsos, H. J. Escalante, L. Sigal, A. Argyros, C. Sminchisescu, R. Bowden, and S. Sclaroff. 2013a. ChaLearn multi-modal gesture recognition 2013: Grand challenge and workshop summary. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*, pp. 365–368.

S. Escalera, J. Gonzàlez, X. Baró, M. Reyes, O. Lopes, I. Guyon, V. Athistos, and H. J. Escalante. 2013b. Multi-modal gesture recognition challenge 2013: Dataset and results. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*, pp. 445–452.

S. Escalera, X. Baró, J. Gonzàlez, M. A. Bautista, M. Madadi, M. Reyes, V. Ponce-López, H. J. Escalante, J. Shotton, and I. Guyon. 2015. ChaLearn looking at people challenge 2014: Dataset and results. In L. Agapito, M. M. Bronstein, and C. Rother, eds., *Computer Vision – ECCV 2014 Workshops*, volume LNCS 8925, pp. 459–473. Springer International Publishing, Switzerland.

S. Escalera, V. Athitsos, and I. Guyon. 2016. Challenges in multimodal gesture recognition. *Journal of Machine Learning Research*, 17: 1–54.

J. Foote. 1999. An overview of audio information retrieval. *Multimedia Systems*, 7(1): 2–10.

S. Goldin-Meadow and M. W. Alibali. 2013. Gesture's role in speaking, learning, and creating language. *Annual Review of Psychology*, 64: 257–283.

S. Goldin-Meadow, D. Wein, and C. Chang. 1992. Assessing knowledge through gesture: Using children's hands to read their minds. *Cognition and Instruction*, 9(3): 201–219.

S. Goldin-Meadow, H. Nusbaum, S. D. Kelly, and S. Wagner. 2001. Explaining math: gesturing lightens the load. *Psychological Science*, 12(6): 516–522.

U. Hadar, A. Burstein, R. Krauss, and N. Soroker. 1998a. Ideational gestures and speech in brain-damaged subjects. *Language and Cognitive Processes*, 13(1): 59–76.

U. Hadar, D. Wenkert-Olenik, R. Krauss, and N. Soroker. 1998b. Gesture and the processing of speech: neuropsychological evidence. *Brain and Language*, 62(1): 107–126.

A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng. 2014. Deep speech: Scaling up end-to-end speech recognition. *Computing Research Repository*, arXiv:1412.5567v2.

H. Hermansky, D. P. W. Ellis, and S. Sharma. 2000. Tandem connectionist feature extraction for conventional HMM systems. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 3, pp. 1635–1638.

G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6): 82–97.

A. B. Hostetter. 2011. When do gestures communicate? A meta-analysis. *Psychological Bulletin*, 137(2): 297–315.

HTK, 2009. http://htk.eng.cam.ac.uk/. (last accessed January 6, 2017).

X. Huang, S. Oviatt, and R. Lunsford. 2006. Combining user modeling and machine learning to predict users' multimodal integration patterns. In S. Renals, S. Bengio, and J. G. Fiscus, eds., *Machine Learning for Multimodal Interaction*, volume LNCS 4299, pp. 50–62. Springer-Verlag, Berlin, Germany.

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

S. Ji, W. Xu, M. Yang, and K. Yu. 2013. 3D convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1): 221–231.

M. Johnston. 1998. Unification-based multimodal parsing. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Conference on Computational Linguistics (ACL-COLING)*, volume 1, pp. 624–630.

M. Johnston and S. Bangalore. 2000. Finite-state multimodal parsing and understanding. In *Proceedings of the Conference on Computational Linguistics (COLING)*, volume 1, pp. 369–375.

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

G. Keren, A. El-Desoky Mousa, O. Pietquin, S. Zafeiriou, and B. Schuller. 2017. Deep learning for multisensorial and multimodal interaction. In S. Oviatt, B. Schuller, P. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, eds., *The Handbook of Multimodal-Multisensor Interfaces, Volume 2: Signal Processing, Architectures, and Detection of Emotion and Cognition*. Morgan Claypool Publishers, San Rafael, CA.

Kinect, 2016. https://developer.microsoft.com/en-us/windows/kinect. (last accessed January 6, 2017).

S. Kopp. 2013. Giving interaction a hand: deep models of co-speech gesture in multimodal systems. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*, pp. 245–246.

S. Kopp and K. Bergmann. 2017. Using cognitive models to understand multimodal processes: The case for speech and gesture production. In S. Oviatt, B. Schuller, P. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, eds., *The Handbook of Multimodal-Multisensor Interfaces, Volume 1: Foundations, User Modeling, and Multimodal Combinations*. Morgan Claypool Publishers, San Rafael, CA.

N. Krahnstoever, S. Kettebekov, M. Yeasin, and R. Sharma. 2002. A real-time framework for natural multimodal interaction with large screen displays. In *Proceedings of the International Conference on Multimodal Interfaces (ICMI)*, pp. 349–354.

R. M. Krauss. 1998. Why do we gesture when we speak? *Current Directions in Psychological Science*, 7(2): 54–60.

R. M. Krauss, Y. Chen, and R. F. Gottesman. 2000. Lexical gestures and lexical access: a process model. In D. McNeill, ed., *Language and Gesture*, pp. 261–283. Cambridge University Press, Cambridge, England.

D. Lalanne, L. Nigay, P. Palanque, P. Robinson, J. Vanderdonckt, and J.-F. Ladry. 2009. Fusion engines for multimodal input: a survey. In *Proceedings of the International Conference on Multimodal Interfaces (ICMI-MLMI)*, pp. 153–160.

I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. 2008. Learning realistic human actions from movies. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8.

J. Liu and M. Kavakli. 2010. A survey of speech-hand gesture recognition for the development of multimodal interfaces in computer games. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1564–1569.

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

P. Maragos, V. Pitsikalis, A. Katsamanis, G. Pavlakos, and S. Theodorakis. 2016. On shape recognition and language. In M. Breuß, A. Bruckstein, P. Maragos, and S. Wuhrer, eds., *Perspectives in Shape Analysis*, pp. 321–344. Springer International Publishing, Switzerland.

D. McNeill. 1985. So you think gestures are nonverbal? *Psychological Review*, 92(3): 350–371.

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

D. McNeill. 2005. *Gesture and Thought*. The University of Chicago Press, Chicago, IL.

D. McNeill, S. D. Duncan, J. Cole, S. Gallagher, and B. Bertenthal. 2008. Growth points from the very beginning. *Interaction Studies*, 9(1): 117–132.

M. Miki, N. Kitaoka, C. Miyajima, T. Nishino, and K. Takeda. 2014. 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): 2:1–2:7.

S. Mitra and T. Acharya. 2007. Gesture recognition: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 37(3): 311–324.

A.-r. Mohamed, G. Dahl, and G. Hinton. 2009. Deep Belief Networks for phone recognition. In *Proceedings of the NIPS Workshop on Deep Learning for Speech Recognition and Related Applications*.

C. Monnier, S. German, and A. Ost. 2015. A multi-scale boosted detector for efficient and robust gesture recognition. In L. Agapito, M. M. Bronstein, and C. Rother, eds., *Computer Vision – ECCV 2014 Workshops*, volume LNCS 8925, pp. 491–502. Springer International Publishing, Switzerland.

L.-P. Morency, A. Quattoni, and T. Darrell. 2007. Latent-dynamic discriminative models for continuous gesture recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8.

N. Morgan and H. Bourlard. 1990. Continuous speech recognition using multilayer perceptrons with hidden Markov models. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pp. 413–416.

P. Morrel-Samuels and R. M. Krauss. 1992. Word familiarity predicts temporal asynchrony of hand gestures and speech. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(3): 615–622.

E. Morsella and R. M. Krauss. 2004. The role of gestures in spatial working memory and speech. *The American Journal of Psychology*, 117(3): 411–424.

C. Müller, A. Cienki, E. Fricke, S. H. Ladewig, D. McNeill, and S. Teßendorf, eds. 2013. *Body – Language – Communication: An International Handbook on Multimodality in Human Interaction*, volume HSK 38.1. De Gruyter Mouton, Berlin, Germany.

K. Nandakumar, K. W. Wan, S. M. A. Chan, W. Z. T. Ng, J. G. Wang, and W. Y. Yau. 2013. A multi-modal gesture recognition system using audio, video, and skeletal joint data. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*, pp. 475–482.

N. Neverova, C. Wolf, G. Paci, G. Sommavilla, G. W. Taylor, and F. Nebout. 2013. A multi-scale approach to gesture detection and recognition. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, pp. 484–491.

N. Neverova, C. Wolf, G. W. Taylor, and F. Nebout. 2015. Multi-scale deep learning for gesture detection and localization. In L. Agapito, M. M. Bronstein, and C. Rother, eds., *Computer Vision – ECCV 2014 Workshops*, volume LNCS 8925, pp. 474–490. Springer International Publishing, Switzerland.

J. Nielsen. 1986. A virtual protocol model for computer-human interaction. *International Journal of Man-Machine Studies*, 24(3): 301–312.

L. Nigay and J. Coutaz. 1993. A design space for multimodal systems: concurrent processing and data fusion. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pp. 172–178.

S. Nobe. 2000. Where do most spontaneous representational gestures actually occur with respect to speech. In D. McNeill, ed., *Language and Gesture*, pp. 186–198. Cambridge University Press, Cambridge, England.

S. Oviatt and P. R. Cohen. 2015. *The Paradigm Shift to Multimodality in Contemporary Computer Interfaces*. Morgan Claypool Publishers, San Rafael, CA.

G. Pavlakos, S. Theodorakis, V. Pitsikalis, A. Katsamanis, and P. Maragos. 2014. Kinect-based multimodal gesture recognition using a two-pass fusion scheme. In *Proceedings of the International Conference on Image Processing (ICIP)*, pp. 1495–1499.

V. I. Pavlovic, R. Sharma, and T. S. Huang. 1997. Visual interpretation of hand gestures for human-computer interaction: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7): 677–695.

X. Peng, L. Wang, Z. Cai, and Y. Qiao. 2015. Action and gesture temporal spotting with super vector representation. In L. Agapito, M. M. Bronstein, and C. Rother, eds., *Computer Vision – ECCV 2014 Workshops*, volume LNCS 8925, pp. 518–527. Springer International Publishing, Switzerland.

L. Pigou, S. Dieleman, P.-J. Kindermans, and B. Schrauwen. 2015. Sign language recognition using convolutional neural networks. In L. Agapito, M. M. Bronstein, and C. Rother, eds., *Computer Vision*

– *ECCV 2014 Workshops*, volume LNCS 8925, pp. 572–578. Springer International Publishing, Switzerland.

K. J. Pine, H. Bird, and E. Kirk. 2007. The effects of prohibiting gestures on children's lexical retrieval ability. *Developmental Science*, 10(6): 747–754.

P. K. Pisharady and M. Saerbeck. 2015. Recent methods and databases in vision-based hand gesture recognition: A review. *Computer Vision and Image Understanding*, 141: 152–165.

V. Pitsikalis, A. Katsamanis, S. Theodorakis, and P. Maragos. 2015. Multimodal gesture recognition via multiple hypotheses rescoring. *Journal of Machine Learning Research*, 16: 255–284.

G. Potamianos, E. Marcheret, Y. Mroueh, V. Goel, A. Koumbaroulis, A. Vartholomaios, and S. Thermos. 2017. Audio and visual modality combination in speech processing applications. In S. Oviatt, B. Schuller, P. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, eds., *The Handbook of Multimodal-Multisensor Interfaces, Volume 1: Foundations, User Modeling, and Multimodal Combinations*. Morgan Claypool Publishers, San Rafael, CA.

L. Rabiner and B.-H. Juang. 1993. *Fundamentals of Speech Recognition*. Prentice Hall, Englewood Cliffs, NJ.

L. R. Rabiner. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2): 257–286.

F. H. Rauscher, R. M. Krauss, and Y. Chen. 1996. Gesture, speech, and lexical access: The role of lexical movements in speech production. *Psychological Science*, 7(4): 226–231.

S. S. Rautaray and A. Agrawal. 2015. Vision based hand gesture recognition for human computer interaction: a survey. *Artificial Intelligence Review*, 43(1): 1–54.

B. Rimé. 1982. The elimination of visible behaviour from social interactions: Effects on verbal, nonverbal and interpersonal variables. *European Journal of Social Psychology*, 12(2): 113–129.

I. Rodomagoulakis, N. Kardaris, V. Pitsikalis, E. Mavroudi, A. Katsamanis, A. Tsiami, and P. Maragos. 2016. Multimodal human action recognition in assistive human-robot interaction. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2702–2706.

M. Rose and J. Douglas. 2001. The differential facilitatory effects of gesture and visualisation processes on object naming in aphasia. *Aphasiology*, 15(10-11): 977–990.

R. C. Rose. 1992. Discriminant wordspotting techniques for rejecting non-vocabulary utterances in unconstrained speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pp. 105–108.

R. C. Rose and D. B. Paul. 1990. A hidden Markov model based keyword recognition system. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pp. 129–132.

G. Saon, T. Sercu, S. J. Rennie, and H. J. Kuo. 2016. The IBM 2016 English conversational telephone speech recognition system. *Computing Research Repository*, arXiv/1604.08242v2.

R. Schwartz and S. Austin. 1991. A comparison of several approximate algorithms for finding multiple (N-best) sentence hypotheses. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pp. 701–704.

A. Shah. 2012. Use voice, gestures to control TV. *PCWorld Magazine*. http://www.pcworld.com/article/253223/use_voice_gestures_to_control_tv.html.

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

M. A. Singer and S. Goldin-Meadow. 2005. Children learn when their teacher's gestures and speech differ. *Psychological Science*, 16(2): 85–89.

G. W. Taylor, R. Fergus, Y. LeCun, and C. Bregler. 2010. Convolutional learning of spatio-temporal features. In K. Daniilidis, P. Maragos, and N. Paragios, eds., *Computer Vision – ECCV 2010 (Part VI)*, volume LNCS 6316, pp. 140–153. Springer-Verlag, Berlin, Germany.

L. A. Thompson and D. W. Massaro. 1986. Evaluation and integration of speech and pointing gestures during referential understanding. *Journal of Experimental Child Psychology*, 42(1): 144–168.

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

A. Vedaldi and B. Fulkerson, 2008. VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/. (last accessed January 6, 2017).

C. Vogler and D. Metaxas. 2001. A framework for recognizing the simultaneous aspects of American sign language. *Computer Vision and Image Understanding*, 81(3): 358–384.

P. Wagner, Z. Malisz, and S. Kopp. 2014. Gesture and speech in interaction: An overview. *Speech Communication*, 57: 209–232.

S. B. Wang, A. Quattoni, L.-P. Morency, D. Demirdjian, and T. Darrell. 2006. Hidden conditional random fields for gesture recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pp. 1521–1527.

G. Willems, T. Tuytelaars, and L. Van Gool. 2008. An efficient dense and scale-invariant spatio-temporal interest point detector. In D. Forsyth, P. Torr, and A. Zisserman, eds., *Computer Vision – ECCV 2008 (Part II)*, volume LNCS 5303, pp. 650–663. Springer-Verlag, Berlin, Germany.

J. G. Wilpon, L. R. Rabiner, C.-H. Lee, and E. R. Goldman. 1990. Automatic recognition of keywords in unconstrained speech using hidden Markov models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(11): 1870–1878.

A. D. Wilson and A. F. Bobick. 1999. Parametric hidden Markov models for gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(9): 884–900.

D. Wu, L. Pigou, P.-J. Kindermans, N. Le, L. Shao, J. Dambre, and J.-M. Odobez. 2016. Deep dynamic neural networks for multimodal gesture segmentation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8): 1583–1597.

J. Wu and J. Cheng. 2014. Bayesian co-boosting for multi-modal gesture recognition. *Journal of Machine Learning Research*, 15: 3013–3036.

J. Wu, J. Cheng, C. Zhao, and H. Lu. 2013. Fusing multi-modal features for gesture recognition. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*, pp. 453–460.

L. Wu, S. L. Oviatt, and P. R. Cohen. 1999. Multimodal integration-a statistical view. *IEEE Transactions on Multimedia*, 1(4): 334–341.

W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig. 2017. The Microsoft 2016 conversational speech recognition system. In *Proceedings of the IEEE International*

Conference on Acoustics, Speech and Signal Processing (ICASSP).

A. Yao, J. Gall, G. Fanelli, and L. Van Gool. 2011. Does human action recognition benefit from pose estimation? In *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 67.1–67.11.

S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. 2002. *The HTK Book*. Cambridge University Engineering Department, Cambridge, England.

D. Yu and L. Deng. 2011. Deep learning and its applications to signal and information processing [exploratory DSP]. *IEEE Signal Processing Magazine*, 28(1): 145–154.

M. Zammit and G. Schafer. 2011. Maternal label and gesture use affects acquisition of specific object names. *Journal of Child Language*, 38(1): 201–221.

T. Zhang and B. Yu. 2005. Boosting with early stopping: Convergence and consistency. *The Annals of Statistics*, 33(4): 1538–1579.

P. Zukow-Goldring. 1996. Sensitive caregiving fosters the comprehension of speech: When gestures speak louder than words. *Early Development and Parenting*, 5(4): 195–211.