Chapter 15
On Shape Recognition and Language

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

Abstract Shapes convey meaning. Language is efficient in expressing and structuring meaning. The main thesis of this chapter is that by integrating shape with linguistic information shape recognition can be improved in performance. It broadens the concept of shape to visual shapes that include both geometric and optical information and explores ways that additional linguistic information may help with shape recognition. Towards this goal, it briefly describes some shape categories which have the potential of better recognition via language, with emphasis on gestures and moving shapes of sign language, as well as on cross-modal relations between vision and language in videos. It also draws inspiration from psychological studies that explore connections between gestures and human languages. Afterwards, it focuses on the broad class of multimodal gestures that combine spatio-temporal visual shapes with audio information. In this area, an approach is reviewed that significantly improves multimodal gesture recognition by fusing 3D shape information from motion-position of gesturing hands/arms and spatio-temporal handshapes in color and depth visual channels with audio information in the form of acoustically recognized sequences of gesture words.

15.1 Introduction

This chapter explores the fusion of shape and linguistic information for improving shape recognition. While its main objective is to address the computer vision problem of shape recognition for shape categories where linguistic information is available by using statistical pattern classification methodologies, it also draws inspiration from psychological studies that explore connections between shapes and human languages. Towards this goal, we broaden the meaning of "shape" to

P. Maragos (✉) • V. Pitsikalis • A. Katsamanis • S. Theodorakis
School of Electrical and Computer Engineering, National Technical University of Athens, Athens 15773, Greece
e-mail: maragos@cs.ntua.gr; vpitsik@cs.ntua.gr; nkatsam@cs.ntua.gr; sth@cs.ntua.gr

G. Pavlakos
Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104, USA
e-mail: pavlakos@seas.upenn.edu

M. Breuß et al. (eds.), *Perspectives in Shape Analysis*, Mathematics and Visualization, DOI 10.1007/978-3-319-24726-7_15

include not only geometric but additional optical attributes; this augmented shape information is referred to as “visual shape”. Thus, as explained in Sect. 15.2.1, shape is meant here in a broader sense of visual information that may encompass brightness, color, depth and time dynamics, even if the main channel is the 2D geometrical shape (or the projection-silhouette of a 3D shape) as time evolves. It focuses on gesture shapes, inspired by the long-term studies of the importance of gestures for the origins of human language and their synergy with speech [3, 25, 31, 47].

We begin with Sect. 15.2 that clarifies how we mean the information conveyed by a shape and in which ways it can be supplemented by linguistic information. We also use statistical inference to intuitively explain how shape recognition may benefit from additional linguistic information. Next, Sect. 15.3 provides a brief survey of shape categories which have the potential of better recognition by combining visual with linguistic information, with emphasis on gestures and moving shapes of sign language, as well as on cross-modal relations between vision and language in videos. This is followed by a motivating Sect. 15.4 on the importance of gestures for human communication. Afterwards, we focus in Sect. 15.5 on the main paradigm of the chapter, which is the broad class of multimodal gestures combining spatio-temporal shapes and other visual cues with audio information in the form of sequences of spoken commands accompanying the gestures; in this section we review an approach [37, 38] that fuses shapes with linguistic information, which is audio-visually expressed, for significantly improving the automated recognition of multimodal gestures. While discussing the examples of both Sects. 15.3 and 15.5 we draw analogies with the main ideas of this chapter.

15.2 Visual Shapes and Linguistic Information

15.2.1 Visual Shapes

Shapes are traditionally perceived and understood as objects of geometry, two-dimensional (2D) or three-dimensional (3D). For a better understanding, perception attributes may be added to them, e.g. as in Gestalt psychology. For automated shape recognition, the computer vision community further explores broader appearance characteristics of shapes by viewing them (whenever possible) as gray intensity images that have both shape and texture. Thus, if 2D shapes are perceived from images, they obtain a third dimension of brightness texture. Instead of brightness, we may also add color to a 2D shape. Temporal dynamics are also important in recognizing moving shapes. Another way of adding a third dimension to a 2D shape to be recognized as a projected silhouette of a 3D object is by using depth. For 3D shapes, including brightness or time evolution will add a fourth dimension.

Thus, in addition to their 2D main projection or silhouette, shapes of world objects can have some additional geometric attributes such as depth and region summary as exemplified by their skeleton axis and its branch points, or even *optical attributes*, e.g. intensity, color, as well as motion (in case of moving shapes) possibly represented by dynamics of the above attributes as time evolves. We shall call this augmented shape information a **visual shape**, meaning that it contains attributes both from geometry (2D or 3D) and optics (photometry and motion). A rich category of such visual shapes that include all the above attributes and will be the main paradigm of this chapter are *gestures*. Both from a human perception and a computer representation viewpoint, gestures comprise several information streams which include a main 2D shape information such as the projection of the handshapes and possibly the moving arms on the image plane, color information, 3D shape, 3D motion, and by using appropriate sensors or computer vision algorithms they can be supplemented with depth and skeleton information. This is illustrated in Fig. 15.1 through an example showing a user performing the Italian gesture "basta" ("that's enough!"). We sampled the video of a user performing this gesture and selected non-uniformly five frames to depict the most important states of the "basta" gesture as time evolves. We supplement the RGB frames with skeleton and depth information, as well as images of the right or both handshapes. In this example the RGB, depth

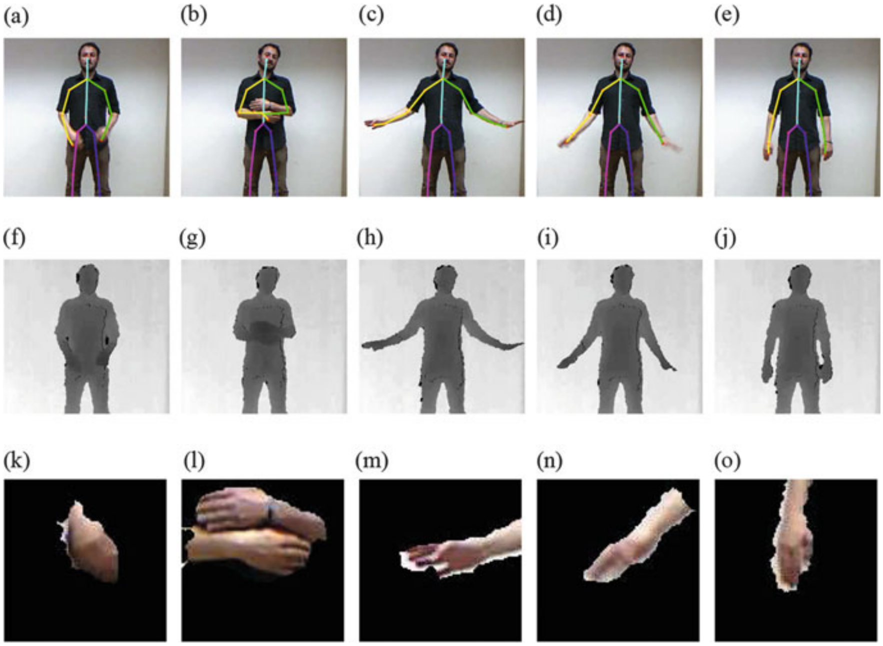


Fig. 15.1 Sequence of frames sampled for a video of a user performing the Italian gesture "basta" ("that's enough!"), obtained with a Kinect sensor. Each column corresponds to a different temporal section of the gesture performance, covering the overall range of motion. Given the start and the end frames here, the duration of the gesture is 36 frames, i.e. 1.8 s (with the frame rate at 20 fps). *First row*: RGB frames accompanied with the skeleton of the user that is superimposed on them. *Second row*: The respective depth frames. *Third row*: Images of the segmented handshapes

and skeleton data were provided by a Kinect sensor. The skeleton information of this sensor includes the human skeleton axis and its branch points, such as hands' centers, elbows, shoulders, face, knees and other critical points.

As exemplified in Fig.15.1, shape information can become richer, and hence its recognition easier, if we augment the geometry of a shape with optical attributes. We offer an intuitive explanation from the domain of statistical pattern classification, by using Bayesian inference. Let S_i represent the i-th class from a collection of shape classes. Suppose we are given measurable data $\mathscr{D}$, which may contain either only geometric information G or geometric and optical information O, and the goal is to infer the shape class given the data via the maximum-a-posteriori principle. If we have information only from geometry, then $\mathscr{D} = \{G\}$, and

$$P(S_i/\mathscr{D}) = \frac{P(S_i)P(G/S_i)}{P(\mathscr{D})} \tag{15.1}$$

where $P(\cdot)$ denotes probability or likelihood. In the case of geometry plus optics, $\mathscr{D} = (G, O)$ and hence

$$P(S_i/\mathscr{D}) = \frac{P(S_i)P(G/S_i)P(O/S_i, G)}{P(\mathscr{D})} \tag{15.2}$$

In the above combined case, the deciding numerator of the right hand side, excluding the prior class probability $P(S_i)$ which is common in both cases (15.1) and (15.2), is a product of two terms, the probability of optical data given the shape class and geometry times the probability of the geometric data given the shape class. By exploiting these two terms we may be able to increase the discriminatory potential of their product. Thus, we may improve the classification of the shape by using statistical knowledge about both its corresponding geometric and optical data, whenever such information is available.

From the domain of philosophy, an extreme such example of the richness of visual shapes versus geometric shapes is Plato's allegory of the cave (presented in his work "The Republic") where silhouettes of real world objects, whose fire-produced shadows are cast on a cave wall while the real objects are being moved behind human spectators, cannot be recognized. In contrast, if the same real objects are seen with direct eye contact and under the sunlight, they reveal their true identity. In Fig. 15.2 we attempted to create an example that illustrates only the visual aspects of the cave allegory. Namely, Fig. 15.2 shows time snapshots from a video of people running and contrasts the complete visual perception provided by the video RGB frames (shape geometry plus color) versus the obviously poorer insufficient information of the 2D silhouettes (shape geometry only) of the moving objects. As *motion* played an important role in the previous gesture sequence of Fig.15.1, we also see in Fig. 15.2 that motion is an important visual cue for understanding of moving shapes.

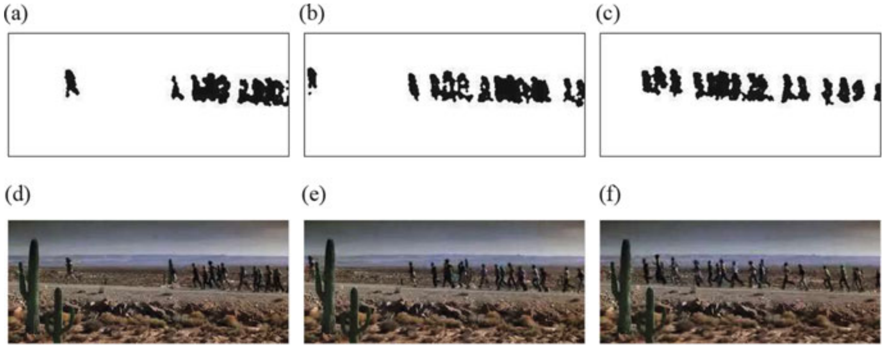


Fig. 15.2 Moving shapes: time sequence of sample frames from a video showing people running (*bottom row*) and their silhouettes (*top row*). The frame rate for this video is at 30 fps. Second frame column is apart from the first by 34 frames (1.13 s), while third frame column is apart from the second by 48 frames (1.6 s)

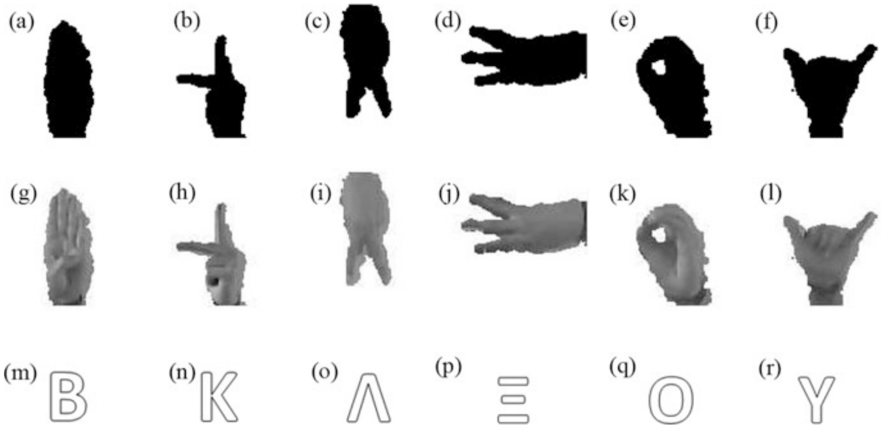


Fig. 15.3 Greek sign language alphabet: shapes, images, letters

So far, one main conclusion is that a *geometric shape*, defined as a 2D or 3D set of points representing an object in the Euclidean space, is a minimalistic form of a *visual shape*, where "visual" means augmenting geometry with optical attributes.

15.2.2 Adding Linguistic Information

The main message from the previous discussion, i.e. that shape inference is enriched if we couple geometry with optics, is further illustrated in the first two rows of Fig. 15.3. The top row shows only silhouettes of handshapes from a sign language. The silhouette only information has some ambiguities, one of which is the question whether the front or back side of the handshapes is visible. In contrast, the middle row shows their corresponding gray images (shape plus brightness texture), i.e. an example of what we call visual shapes, which disambiguate both the visible side of the handshape and add texture details on the visible surface. If a viewer did not know sign language, the first two rows of Fig. 15.3 would just be some handshapes with shape differences among them. However, if we add the information of the third row which corresponds these handshapes with distinct letters of the Greek sign language alphabet, then we have augmented information of a visual shape plus language. This addition of linguistic information can improve the recognition of such handshapes, both from an intuitive viewpoint and from a Bayesian inference viewpoint. To detail the latter (as inspired from statistical speech recognition [23, 40]), assume for example that we are given a time sequence $S = (s_1, s_2, \dots, s_T)$ of shapes s_i from a visual language, in the form of spatio-temporal visual data, and each shape corresponds to a word w_i, then we can recognize the unknown sequence of visual words $W = (w_1, w_2, \dots, w_T)$ by estimating it via the maximum-a-posteriori principle:

$$W^* = \arg\max_W P(W/S) = \arg\max_W \frac{P(S/W)P(W)}{P(S)} \tag{15.3}$$

Thus, the likelihood $P(S/W)$ of the visual shape sequence given its linguistic structure is combined with the prior probability $P(W)$ of the linguistic sequence; this can potentially improve the recognition by exploiting statistical knowledge of the language, e.g. if the n-gram probabilities $P(w_i/w_{i-1} \cdots w_{i-n+1})$ are known.

One way of creating a correspondence between visual shapes and words of some language is via *clustering*. As further elaborated in Sect. 15.4 on the importance of gestures for human communication, imagine given a sequence of visual shape data that span a domain of visual realizations of concepts or objects common to some human community and are represented by visual feature vectors. Then, by some clustering method such as for example the K-means algorithm we can partition the data over this domain into cells (which are regions of the feature space), each representing a concept or object. The mapping of visual shapes in each cell to the cell centroid is some form of feature encoding known as *vector quantization*. Then, these centroids can play the role of words or subword units in some language. In addition to its general usefulness in pattern recognition and machine learning [5, 14, 46], clustering via vector quantization has also been used in signal processing for data compression [20], in speech recognition for converting continuous feature vectors into discrete patterns [40], and in computer vision for action or object recognition based on the *bag of visual words* approach [19, 27, 43].

In the following sections we shall briefly describe some paradigms where visual shape information is supplemented by additional linguistic information. We distinguish three cases:

(1) Relationships between visual shapes and linguistic information. These include (i) direct correspondences as for example in Fig. 15.3 and the pictograms

mentioned in the beginning of Sect. 15.3; (ii) cross-modal relationships in Sect. 15.3.2 between visual objects, represented by their shape information, and linguistic information as corresponding words in text or related audio sounds, employed in a multimedia analysis framework.

(2) In the second case, we employ linguistic information from sign language at the level of visual phonetics. For example, in sign language recognition (Sect. 15.3.1) the video segment corresponding to the visual word of an isolated sign is decomposed into a time sequence of subunits that have a phonetic meaning.

(3) In the third case, which focuses on multimodal gesture recognition (Sect. 15.5), linguistic information is expressed in parallel audio and visual modalities: in the visual stream, gestures occur in a time sequence; in parallel, in the audio stream a sequence of corresponding keywords (or spoken commands) accompanies the visual gestures and provides additional linguistic information.

In all the above paradigms the linguistic information we employ stays only in the specific examples as case studies, and at the level of words or word-subunits; for instance, we do not discuss linguistic structure at the level of sentences.

15.3 Shape and Language Paradigms

Among the earliest paradigms of correspondences between shape and language are the ideographic and logographic writing systems. In the ideographic system the graphemes are the ideograms which are graphic symbols expressing pictorially some concept, independently of any specific language but often assuming some prior convention. A special case are the *pictograms* which further provide a pictorial resemblance with a physical object. Thus, in pictograms there is a direct connection between shape and language. The logographic system is based on logograms which are graphemes that represent words or morphemes and may also contain phonetic elements. Examples of logograms include numerous Egyptian hieroglyphs and Chinese characters. A famous example that may fit in one of the above cases are the shapes on the Phaistos Disk, which was discovered in 1908 at the Minoan palace of Phaistos on the Greek island of Crete, possibly dating from the 2nd millennium B.C.; see Fig. 15.4. Although the ancient Egyptian hieroglyphs have been deciphered after the discovery of the Rosetta Stone in 1799, the glyphs on the Phaistos Disk still remain an archeological mystery.

In the two following subsections we highlight some ideas relevant to this chapter from two broad categories of moving shapes where we encounter numerous correspondences between shapes and language: (i) sign gestures and facial expressions encountered in sign language and (ii) multimodal relationships between vision plus language (audio or text) that are abundant in movie videos.

Fig. 15.4 Phaistos disk (At the archaeological museum of Heraklion, Crete)

15.3.1 Sign Language

Human languages include both spoken and sign languages. Sign languages are natural languages communicable purely by vision via sequences of time-varying 3D shapes. They serve for communication in the Deaf communities, as well as among deaf and hearing people if the latter learn to sign. They convey information and meaning via spatio-temporal visual patterns, which are formed by manual (handshapes) and non-manual cues (facial expressions and upper body motion). A coarse correspondence of a word in spoken language is a sign in sign language. See [15, 28] for surveys of linguistic and cognitive aspects of sign language. The area of computer-based processing and recognition of sign videos is also broadly related to vision-based human-computer interaction using gesture recognition [22].

While significant progress exists in the field of automatic sign language recognition from the computer vision and pattern recognition fields, e.g. see [1, 8, 32, 44, 45, 49] and the references therein, it still remains a quite challenging task especially for continuous sign language. In addition to signs having a complex multi-cue 4D space-time structure, the difficulty in their automatic recognition is also due to the large variability with respect to inter-signer or intra-signer variations of signing while expressing the same concept-word. An example exhibiting such variations is shown in Fig. 15.5. This variation is due to various sources: (i) the physiology of each signer and the manner of his/her signing, (ii) the coarticulation – continuous variability that causes multiple pronunciations, and (iii) the existence of multiple pronunciations per se (e.g. from different dialects). Due to the above variability, instead of recognizing each sign as a whole word, a more efficient approach (inspired by speech recognition) is to decompose signs into *subunits*, resembling the phonemes of speech, and recognize them as a specific sequence of subunits by using some statistical model, e.g. via Hidden Markov Models (HMMs). Clearly, the subunits approach performs much better on large vocabularies and continuous

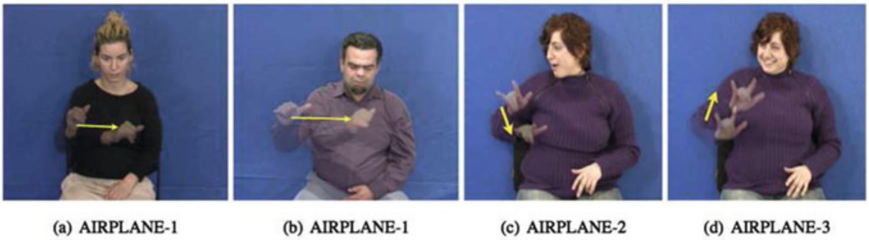


Fig. 15.5 Multiple realizations for sign /airplane/. (**a**) and (**b**) are due to inter-signer variability. (**c**) and (**d**) are due to intra-signer variability. On each image we superimpose the beginning and end frames of the sign with an *arrow*

language; further, the subunits are reusable and help with signer adaptation. In lack of a lexicon, a computational technique to find such subunits is *data-driven*, i.e. perform unsupervised clustering on a large database and use the cluster centroids as subunits. This performs well in several instances, especially when the subunits are pre-classified and statistically modeled based on visual features into dynamic vs. static, as done by Theodorakis et al. [45], where the dynamic or static refers to the type of the signer's hands and arms motion. However, a superior performance accompanied with phonetic interpretability may be obtained if the chosen subunits are also based on the phonetic structure of a sign, as for example by incorporating the Posture-Detention-Transition-Steady Shift (PDTS)[1] system [24] of phonetic labels. A sequence of PDTS phonetic subunits is shown in Fig. 15.6. Pitsikalis et al. [39] combined the phonetic information provided by the PDTS transcriptions of sign videos with the automatically extracted visual features to create (1) statistically trained *phonetic subunits* and a corresponding lexicon, which were then used for (2) optimally aligning (via Viterbi decoding) the data with the phonetic labels and hence providing the missing temporal segmentation, as well as (3) better sign recognition. Thus, we have a clear paradigm of improved shape recognition when the visual information is coupled with linguistic information.

While information and meaning in sign languages are mainly conveyed by moving handshapes, they are also conveyed in part by non-manual cues such as facial expressions. These expressions can be visually modeled by deformable models that encode both geometric shape and brightness texture information. Such a class of models often used in computer vision are the active appearance models (AAMs) [11]. Examples of the deformable geometric masks of such facial AAMs are illustrated in Fig. 15.7, which shows a few frames from a sign sequence that involves eye blinking. The transient phenomenon of eye blinking, where the eyes may take one of the open/closed states, conveys low-level linguistic information such as sentence – and possibly sometimes sign – boundaries, as described in Anton-

[1] In the PDTS system, D is a "hold" but for shorter duration than P. S is a "movement" without acceleration. T is more abrupt motion.

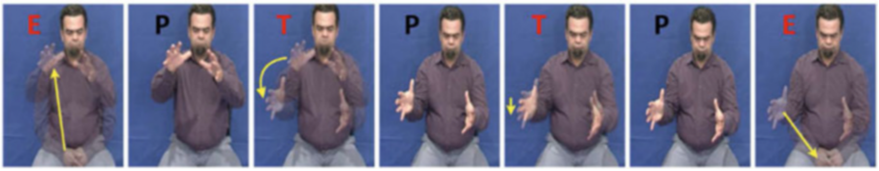


Fig. 15.6 Sample frames from the sign /pile/ from the Greek sign language. Images marked with "T" and "E" represent dynamic segments with the phonetic labels "Transition (T)" and "Epenthesis (E)", visualized by superimposing on the same image the beginning and end frames with an *arrow*. Images marked with "P" represent static segments with the phonetic label "Posture (P)", visualized by a single frame (Figure courtesy of Pitsikalis, Theodorakis, Vogler and Maragos [39])

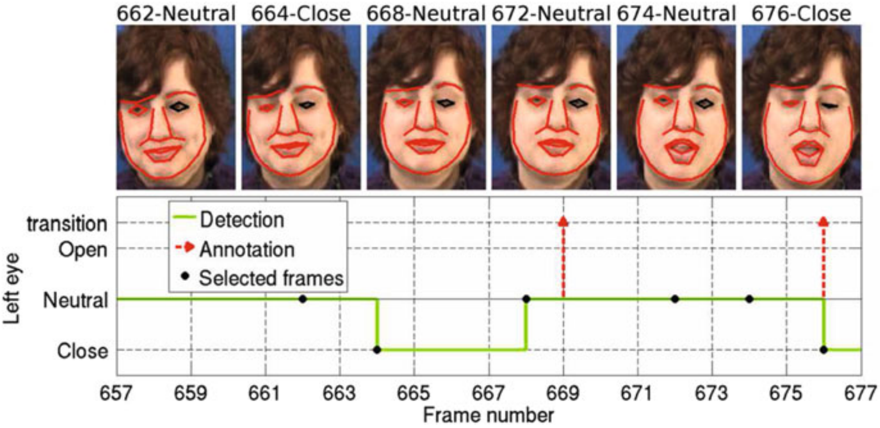


Fig. 15.7 Sign boundary detection based on eye blinking detection on a Greek Sign Language database. Indicative frames (*up*) are marked with a *black dot* in the detection diagram (*down*) (Figure courtesy of Antonakos, Pitsikalis and Maragos [2])

akos et al. [2] and the references therein. The detection of the eye opening/closing transitions can be detected from the changes in the corresponding AAM parameters. Figure 15.7 presents an example of such a detection between neutral-close-neutral (neutral is considered as intermediate) and its correspondence with the annotated sign boundaries. This is another paradigm of synergy between visual shape and language.

15.3.2 Multimodal Relations Between Shapes and Language

Every day communication between people is a blend of different modalities. Humans often combine different pieces of information, e.g. visual and linguistic, in order to communicate and interact. In multimedia data such as multimodal videos, visual, auditory and linguistic information coexist as well. In multimodal videos we encounter a variety of visual objects that we can recognize more easily when there exists a concurrent linguistic reference either in the text domain or as an acoustic event. (Note that linguistic information can also exist in a video without text or audio, e.g. in sign language videos as described in Sect. 15.3.1.) This is one aspect of a broader class of phenomena with audio-visual modality integration, which is an active research area in behavioral psychophysics, e.g. see [48], and in neuroscience where, for instance, brain activity during watching TV programs as measured by fMRI reveals correlations between audio and visual stimuli [7]. From a computational viewpoint, this audio-visual synergy can improve recognition performance in multimedia systems via cross-modal integration, as surveyed in [29] and the references therein. In general, there has been significant evidence that human perception is multimodal and hence perception of visual objects can be improved when different modalities are synergetically employed.

A corpus-based framework for analyzing and modeling multimedia dialectics is the COSMOROE framework [36] which describes the semantic interplay between verbal and non-verbal communication; specifically, the cross-media semantic interrelation between images, language (in the form of either spoken language transcription, graphic/scene text shown on the video, or acoustic stimuli, e.g. human/animal or environmental sounds) and body movement. In Fig. 15.8 we provide two such examples from cross-modal relations between visual shapes and linguistic information. For instance, in a quite complex scene as presented in Fig. 15.8a, where there is interaction between people (with clothing that attracts human attention), the image of the *dog* could go unnoticed; however, the fact that the dog is *barking* guides our look towards it. Same observation applies to Fig. 15.8b as well; the lamp could easily get overlooked if the acoustic stimulus as in the phrase *"Take the lamp out on the porch"* did not take place. This association of a visual object with the linguistic information may render the recognition procedure easier for humans and more robust for computers.

In short, the COSMOROE framework [36] aims at finding and analyzing relations from linguistics to other modalities, especially visual shapes, in multimodal corpora. In parallel, there is a recent trend in computer vision in the opposite

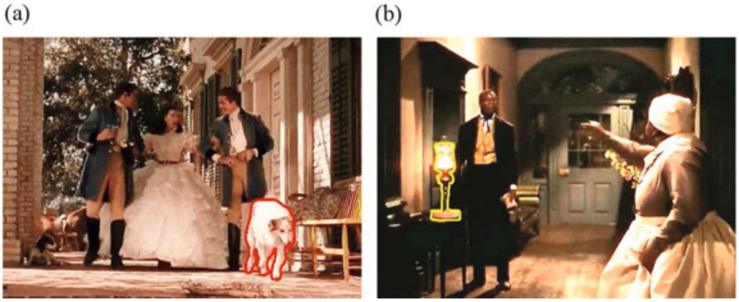


Fig. 15.8 Correspondence between shapes and linguistic information (aural or textual) in movie videos. (**a**) Acoustic event: dog barking. (**b**) Utterance: "Take the lamp out on the porch"

direction, i.e. associating visual objects with linguistic attributes, which can benefit recognitions problems such as action recognition [27] and person recognition [12] in movie videos, as well as general object recognition [18, 35].

15.4 Gestures in Human Communication

In Sect. 15.3.1 on sign language we summarized that certain types of moving bodily shapes can convey linguistic messages that represent complete languages. Here we further extend this idea by providing a brief survey on how gestures have been of great significance in human communication. In particular, according to specific theories [3, 31, 47], they have supported the beginnings of language formation, after which gesture shapes and language can reinforce each other.

By gestures we mean visible actions involving shapes of manual and non-manual bodily motions and postures; most of them are dynamic (i.e. time-varying for part of their duration) and use the hands. Kendon [25] classifies human gestures in (1) Gesticulations, (2) Speech-framed, (3) Pantomimes, (4) Emblems (quotable gestures), and (5) Sign language. The above sequence has been called *Kendon's continuum* [30]. As the numerical index of the gesture class increases, the degree to which speech should accompany a gesture decreases whereas the degree to which a gesture shows language-like properties increases.

The theory that gesture-based human communication evolved first whereas conventional languages evolved later has had many supporters from the antiquity until it became more definite in the eighteenth century; afterwards gesture and sign languages started being studied as natural languages. Wittgenstein in his work [52] on the philosophy of language argued that "What we call meaning must be connected with the primitive language of gestures". In search of the origins of human communication, Tomasello [47] has provided ample evidence about the critical importance of gestures, in particular of the pointing and pantomiming types, for humans to develop (i) social cognitive skills that create a common conceptual ground, including joint attention, shared experience and common cultural knowledge, and (ii) social motives such as requesting, informing, helping and sharing with others. These developments of social cognition and motivation create a *shared intentionality*, as is called by some modern philosophers of action, e.g. [42]. Quoting from [47], "pointing (deictic gestures) direct the attention of a recipient to something in the immediate perceptual environment, whereas pantomiming (iconic gestures) direct the imagination of a recipient to something that typically is not in the immediate perceptual environment by simulating an action, relation, or object". Interestingly apes have also developed pointing (attention-getters) and pantomiming (intention-movements) gestures for their communication. One big difference between the gesture-based ape versus human communication is that for apes it serves individual intentionality, whereas for humans it serves shared intentionality. This shared intentionality is at the heart of the *cooperative model for human communication* [47].

Thus, according to the theory and evidences in [47], the human social cognitive skills and social motivation create a cooperative psychological infrastructure of human communication based on gestures, which laid the foundations for the later development of conventional languages. By “conventional language” we mean a symbolic communicative code, which assumes some preexisting codified form of communication like the gesture modality. Such a linguistic code is based on a non-linguistic infrastructure of intentional understanding and common conceptual ground [47, 51]. From a computational viewpoint, we may conjecture that nowadays, if we are given a collection of gestures referring to a common perceptual ground of objects, then by clustering and feature encoding we could in theory map gestures to some abstract language words which could be the cluster centroids. Of course, after their early development, human conventional languages, mainly spoken languages, evolved into a very creative and versatile form of communication which, despite its complexity, has fundamentally supported and propelled human civilization. In contrast to gestures, the vocal modality in nonhuman mammals remained inflexible and has not created a language. Quoting from [47], “for all mammals, including nonhuman primates, *vocal displays* are mostly unlearned, genetically fixed, emotionally urgent, involuntary, inflexible responses to evolutionarily important events that benefit the vocalizer. In stark contrast, a significant number of nonhuman primate *gestures*, especially those of great apes, are individually learned and flexibly produced communicative acts, involving an understanding of important aspects of individual intentionality.”

Another supporter of the “gesture-first” conjecture is Arbib [3] who supports a theory that human language evolved as a result of biological and cultural evolution starting from simple manual gestures we share with apes, progressing to the imitation of manual skills and pantomime, and culminating to the development of sign language and speech.

In addition to the gesture-first theory which advocates that human language started as non-spoken gestures and signs, there are also combined theories that advocate a fusion of the gesture and speech modality. For example, based on evidence from neurological and psychological data, McNeill [31] argues for a two-phase development of language acquisition in children: The first phase is based only on gestures without speech. Later, when the required brain structures have matured at age about 3–4, the second phase begins and involves both speech and gestures. This *gesture-speech unity* continues in adult life and uniquely characterizes the human language that we have actually evolved as a species.

It is this *multimodal* view of the language, containing both imagery via gestures and linguistic codes via speech, that we further pursue in this chapter by discussing computational approaches to automate its recognition, as explained next in the paradigm of audio-visual gesture recognition.

15.5 Multimodal Gesture Recognition

Multimodal gestures, i.e. time sequences of isolated gestures with simultaneous utterance of the corresponding keyword (or spoken command), is a primary domain where the fusion of visual shapes (gestures) with linguistic information (spoken commands) leads to significantly improved recognition over visual only recognition. They are becoming increasingly useful for human-computer interaction [6, 22, 26, 34]. In this section we highlight the main ideas and method of the chapter authors' recent works in [37] and [38] for the effective recognition of multimodally expressed gestures as performed freely by multiple users. The experiments were performed on a demanding dataset [17] which was acquired via Kinect for the purpose of the ChaLearn multimodal gesture recognition challenge (in conjunction with ACM ICMI 2013) [16]. It comprises multimodal cultural-anthropological gestures of everyday life, in multi-user spontaneous realizations of both spoken and hand-gesture articulations, intermixed with other random and irrelevant hand or body movements and spoken phrases. The use of Kinect enables multimodal capturing and provides four information streams, three visual (RGB color video, depth video, and skeleton with tracking of its branch points) and one aural (audio stream), all essential to multimodal processing. In the next subsections, we briefly review the approach in [37, 38] for multimodal gesture recognition, where the additional employment of speech significantly improves the performance of recognition over using only visual shape information (handshape and skeleton).

15.5.1 Methodology

The multimodal gesture recognition system exploits the color, depth, skeleton and audio signals captured by the Kinect sensor. See Fig. 15.9 for an overall view of the proposed fusion scheme. It extracts features for the handshape configuration, the movement of the hands and the speech signal, and it essentially implements a two-level[2] fusion approach:

> *1st Pass (P1)*: To independently account for the specificities of each of the modalities involved, we first train separate gesture-word models for each modality. These unimodal models are then used to generate a set of possible gesture-word sequence hypotheses for a given recording. Then, this original set of hypotheses is multimodally rescored and resorted.

[2] In the work of [38] the P1/P2 terms are not employed any more compared to [37], since [38] includes several other contributions, the discussion of which is beyond the scope of this chapter. Herein we keep the P1/P2 terms only for descriptive reasons.

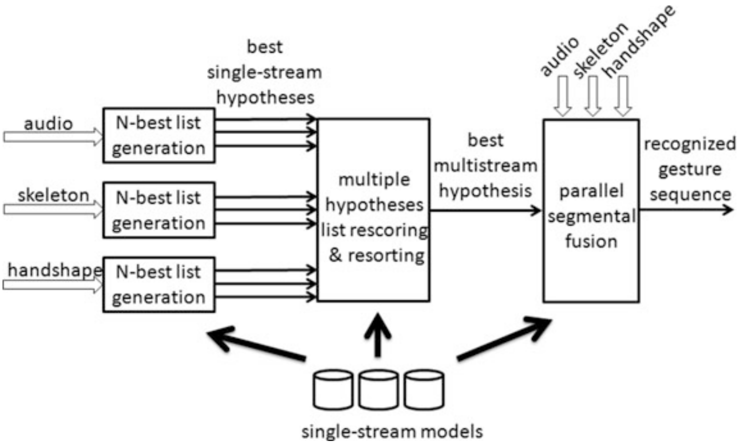


Fig. 15.9 Overview of the multimodal fusion scheme for gesture recognition based on multimodal hypotheses rescoring. Single-stream models are first used to generate possible hypotheses for the observed gesture sequence. The hypotheses are then rescored by all streams and the best one is selected. Finally, the observed sequence is segmented at the temporal boundaries suggested by the selected hypothesis and parallel fusion is applied to classify the resulting segments. Details are given in Sect. 15.5.1.2 (Figure courtesy of Pitsikalis, Katsamanis, Theodorakis and Maragos [38])

2nd Pass (P2): Based on the temporal boundaries of the gestures in the best fused hypothesis, a parallel segmental fusion step as in [49] exploiting all three modalities further improves recognition.

Gestures in our case occur in parallel with their semantically corresponding speech words, without implying however strictly synchronous realizations in all modalities. Given a vocabulary $V = \{g_i\}$, $i = 1, \ldots, |V|$, of multimodal gestures g_i that are to be detected and recognized in a recording and a set $C = \{\mathbf{O}_m\}$, $m = 1, \ldots, |C|$, of measurements from multiple information channels/streams that are concurrently observed, our goal is to generate the best multimodal hypothesis $\mathbf{h}$ for the sequence of gesture appearances, based on these observations. In our experiments, the latter set comprises three streams, namely handshape features, skeleton features and audio spectral features. In essence, any set of information streams can be employed in this framework, although the combination of visual and audio cues significantly enhances recognition results.

15.5.1.1 Single Information Stream Modeling

The modeling methodology essentially follows the keyword-filler paradigm for speech [41, 50] and is based on hidden Markov models (HMMs). For a tutorial on HMMs and their application to speech recognition, the reader is referred to [23, 40]. The problem of recognizing a limited number of gesture-words in a video possibly comprising other heterogeneous events as well, is seen as a keyword detection problem. The gesture-words to be recognized are the keywords and all the rest is ignored. Each gesture-word is modeled by a left-to-right HMM with a common number of states and with Gaussian mixture models (GMMs) representing the state-dependent observation probability distributions. There are also two separate filler HMMs to represent either silence/inactivity, or all other possible events (called "background model" – BM) appearing in that stream.

15.5.1.2 Multimodal Fusion

N-Best Rescoring and Resorting: Using the single stream gesture models and a gesture grammar G, which defines the set of alternative hypotheses allowed, a list of N-best possible hypotheses is initially generated for the unknown sequence for each stream. Specifically, by applying Viterbi decoding [40] we can estimate the best hypothesis $\hat{\mathbf{h}}_m$ per stream:

$$\hat{\mathbf{h}}_m = \arg\max_{\mathbf{h}\in G} \log P(\mathbf{O}_m|\mathbf{h}, \lambda_m), \quad m = 1, \ldots, |C|, \tag{15.4}$$

where $\mathbf{O}_m$ is the observation sequence for modality m, λ_m is the corresponding set of HMM models, and G is the set of alternative hypotheses allowed by the gesture grammar.

Similarly, in the more general case, we can generate a complete list of the N-best gesture-word sequences per stream, and form a set $H = \{\mathbf{h}_1, \ldots, \mathbf{h}_L\}$ of all the hypotheses (L in total) for the available modalities. Given this set, we sort the hypotheses [10, 21, 33] and identify the most likely hypothesis exploiting all modalities. In this direction, we estimate a combined score for each possible gesture sequence as a weighted sum of standardized modality based scores:

$$v_i = \sum_{m=1}^{|C|} w_m v^s_{m,i}, \quad i = 1, \ldots, L \tag{15.5}$$

where the weights w_m for each modality m can be determined experimentally (by maximizing the recognition score on the validation set). The modality-based scores $v^s_{m,i}$ are standardized versions of $v_{m,i}$ which are estimated by means of Viterbi decoding:

$$v_{m,i} = \max_{\mathbf{h}\in G_{h_i}} \log P(\mathbf{O}_m|\mathbf{h}, \lambda_m), \quad i = 1, \ldots, L, \quad m = 1, \ldots, |C|, \tag{15.6}$$

This maximization searches over acceptable gesture sequences that follow a specific hypothesis-dependent finite-state grammar G_{h_i}. Thus, this is a constrained recognition problem where the search space of possible state sequences includes only sequences corresponding to the hypothesis $\mathbf{h}_i$ plus possible variations by keeping the appearances of target gestures unaltered and only allow SIL (silence) and BM (background model) labels to be inserted, deleted and substituted with each other. The most probable gesture-word sequence hypothesis $\mathbf{h}^* = \mathbf{h}_{i^*}$, where $i^* = \arg\max_i v_i$, after this step is the one with the maximum combined score.

Segmental Parallel Step: Herein we exploit the modality-specific time boundaries (found via forced alignment) for the most likely gesture sequence and segment each observation stream, to reduce the recognition problem to a segmental classification one. For every segment and each stream, we compute the log probability:

$$LL^t_{m,j} = \max_{\mathbf{q} \in Q} \log P(\mathbf{O}^t_m, \mathbf{q} | \lambda_{m,j}), \;\; j = 1, \ldots, |V|, \tag{15.7}$$

where t is the time index of the segment, $\lambda_{m,j}$ are the parameters of the HMM model for the gesture g_j and the stream m; $\mathbf{q}$ is a possible state sequence. These segmental scores are linearly combined across modalities to get a multimodal score:

$$LL^t_j = \sum_{m=1}^{|C|} w'_m LL^t_{m,j} \tag{15.8}$$

where w'_m is the stream-weight for modality m set to optimize recognition performance of this step. Finally, the recognized gesture for each segment t is the one with the highest multimodal score. This final stage is expected to give additional improvements, allowing local refinements by exploiting possible benefits of a segmental classification process.

15.5.2 Experimental Results

15.5.2.1 Multimodal Gesture Dataset

For the experimental work we employed the ChaLearn multimodal gesture challenge dataset [17], which focuses on multiple-instance, user-independent learning of gestures from multimodal data. It provides via Kinect RGB and depth images of face and body, user masks, skeleton information, as well as concurrently recorded audio including the speech utterance accompanying the gesture. See top row of Fig. 15.10 for an example of the data. The vocabulary contains 20 Italian cultural-anthropological gestures, performed by 39 users in 13,858 gesture-word instances in total. Gesture recognition over this dataset presents several challenges: presence of distracting gestures, large number of categories, length of gesture sequences, user variety and corresponding variability in gestures and spoken dialects, variations in background and lighting; see Fig. 15.11.

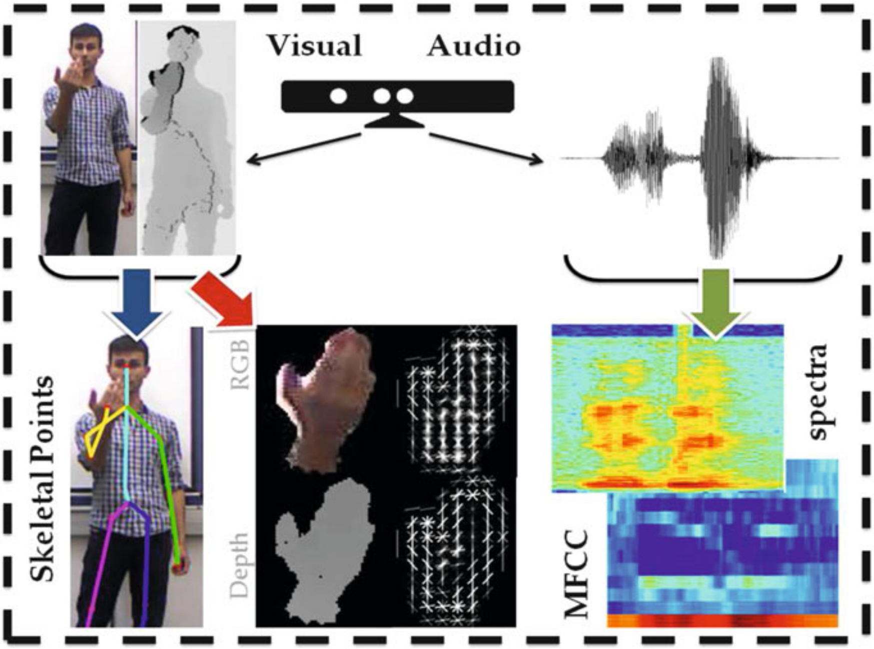


Fig. 15.10 A collection of sample cues as well as extracted features for each modality. *Top row*: visual data (RGB and depth) and audio data. *Bottom row*: visual features (skeletal points, HOGs in the RGB and depth channels) and audio features (MFCCs)

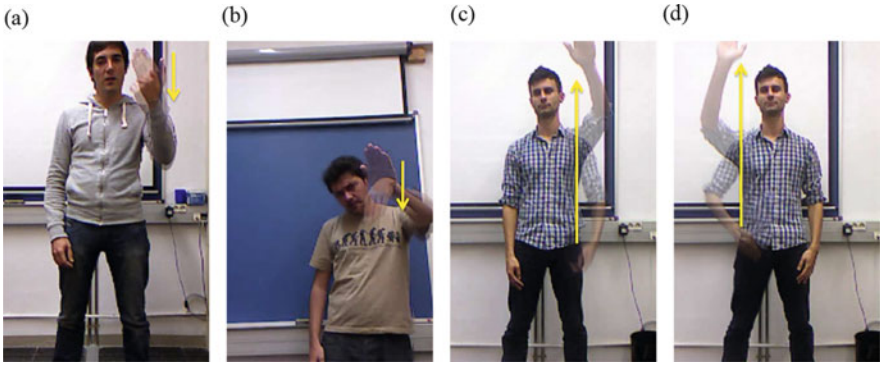


Fig. 15.11 (**a,b**) Arm position variation (low, high) for gesture "vieni qui" ("come here"); (**c,d**) Left- and right-handed instances of gesture "vattene" ("go away"). Gesture motion is visualized by superimposing on the same image the beginning and end frames with an *arrow*

15.5.2.2 Multimodal Features

We statistically train separate HMMs at the level of word-gestures per each modality, i.e. handshape, skeleton and audio.

Table 15.1 Single modalities evaluation expressed as accuracy (in %)

Audio	*Skeleton*	*HandShape*
87.2	49.1	20.2

Table 15.2 Our approach in comparison with the first three places of the ACM 2013 Gesture Challenge

Approach	Accuracy %
Ours [38]	93.3
iva.mm [53]	87.2
wweight [17]	84.6
E.T. [4]	82.9

Handshape Cue: The features employed are Histograms of Oriented Gradients (HOGs) [13] as extracted in both hands' segmented images for both RGB and depth modality. We segment the hands by employing the hand's tracking and by performing threshold depth segmentation. Essentially, any visual descriptor could be computed on the handshape information; HOGs are just an example that is used widely in the literature (e.g. in [9]).

Skeleton Cue: The features employed for the skeleton cue include: the hands' and elbows' 3D position, the hands' 3D position with respect to the corresponding elbow, the 3D direction of the hands' movement, and the 3D distance of hands' centroids.

Audio Cue: To efficiently capture the spectral properties of speech signals, our frontend generates 39 acoustic features every 10 ms. Each feature vector comprises 13 Mel Frequency Cepstral Coefficients (MFCCs) along with their first and second derivatives.

A visualization of the extracted features for all the available modalities is presented in bottom row of Fig. 15.10.

15.5.2.3 Recognition Results

We summarize the most recent[3] experimental results from [38].

In Table 15.1 we show the recognition results for each modality. The results are expressed in *accuracy* (%), which is computed as $100 - \text{WER}$ where WER is the percent word error rate that includes insertions, deletions and substitutions. As observed, the audio modality is the strongest one.

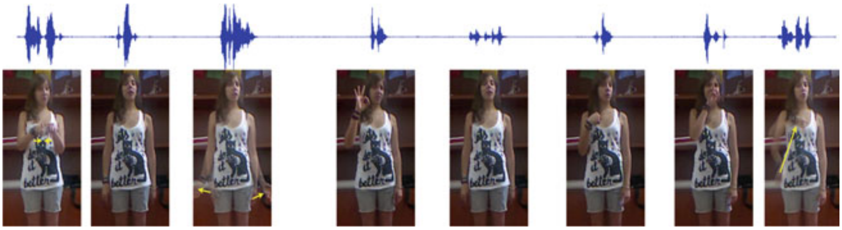
Table 15.2 shows the performance of the proposed multimodal two-pass fusion scheme [38] in comparison with other approaches who participated in the Gesture Challenge [17]. Our scheme begins with a first-pass fusion step (*P1*) leading to

[3] The multimodal gesture recognition system in [38] is an extension of [37], where additional components are included such as voice and gesture activity detection and a gesture-loop grammar, which improve the recognition results.

the best fused hypothesis as a result of the N-best rescoring. Then follows the *P2* component as the second-pass fusion step; in this we employ the gesture-word level segmentation of the above best fused hypothesis, leading to the second-pass fused result and the final recognized words. This multimodal fusion yields a recognition accuracy of 93.3 %, which outperforms the other approaches and reduces the smallest previous error by a relative 47 %.

A gesture sequence decoding example is shown in Fig. 15.12. Herein we illustrate both audio and visual modalities for a word sequence accompanied with the ground truth word-level transcriptions (row: "REF"). In addition we show the decoding output employing the single-audio modality (AUDIO) and the three presented fusion cases ($P1$, $P2$ and $P1 + P2$). As we observe there are several cases where the subject pronounces an out-of-vocabulary (OOV) word and either performs a gesture or not. This indicates the difficulty of the task as these cases should be ignored. By focusing on the recognized word sequence that employs the single-audio modality we notice two insertions (words "PREDERE" and "FAME"). By employing either the $P1$ or $P2$ the above word insertions are corrected as the visual modality is integrated and helps identifying that these segments correspond to OOV words. Further, the single pass fusion components lead to errors which the proposed approach manages to deal with: $P1$ causes insertion of "OK", $P2$ of a word deletion "BM". These are in contrast to $P1 + P2$ which recognizes correctly the whole sentence.

Note that for the above audio-visual fusion on the Gesture Challenge dataset, we implicitly address inter-stream differences, since (a) our modeling deals with not perfectly aligned audio and visual information (we enforce different boundaries for each stream), and (b) with fusion we can handle cases where one stream is less informative than the others. In fact, Fig. 15.12 presents cases (third and sixth frame) where the audio modality is ambiguous (and estimates the wrong word), whereas



REF	DACCORDO	OOV	OOV	OK	OOV	OOV	OOV	SONOSTUFO
AUDIO	DACCORDO	BM	PREDERE	OK	BM	FAME	BM	SONOSTUFO
P1	DACCORDO	BM	BM	OK	BM	BM	OK	SONOSTUFO
P2	DACCORDO	BM	BM	BM	BM	BM	BM	SONOSTUFO
P1+P2	DACCORDO	BM	BM	OK	BM	BM	BM	SONOSTUFO

Fig. 15.12 An example of recognizing a gesture-word sequence. Audio (*top*) and visual modalities (second) via a sequence of images for a word sequence. Ground truth transcriptions ("REF"). Decoding results for the single-audio modality (AUDIO) and the three different fusion schemes (P1, P2 and P1+P2). Errors are highlighted: deletions (*blue* color) and insertions (*green* color). A background model (BM) models the out-of-vocabulary (OOV) words (Figure courtesy of Pitsikalis, Katsamanis, Theodorakis and Maragos [38])

for the visual streams we are more confident about the gesture, so with fusion of the results we get the correct gesture-word for these segments.

15.6 Conclusions

In this chapter we have proposed a broader view of shapes and their temporal sequences as communicative devices. In particular, we have emphasized the connections between shape and language and have argued for improving shape recognition by adjoining linguistic information. To illustrate this idea we have provided several paradigms including examples from sign recognition and shape-language relations in multimodal videos. Then, we have focused on the class of multimodal gesture sequences and showed the great improvement in gesture recognition achievable by fusing visual gesture shapes with spoken commands in multimodal videos. These paradigms employed some specific methodologies from pattern recognition, i.e. HMMs, motivated by the relative success they have had in speech recognition on integrating acoustic with linguistic information, but there are also alternative machine learning approaches that could be applied. However, despite the possibility of employing more efficient methodologies, the main thesis of this chapter remains the capability of improving shape recognition by adding linguistic information. This is possible and meaningful for those categories of shapes whose modeling can be considered in a linguistic context.

Acknowledgements We wish to thank Niki Efthymiou and Nancy Zlatintsi at NTUA CVSP Lab for Fig. 15.8 and discussions related to Sect. 15.3.2. This research work was supported by the project COGNIMUSE which is implemented under the ARISTEIA Action of the Operational Program Education and Lifelong Learning and is co-funded by the European Social Fund and Greek National Resources. It was also partially supported by the European Union under the project MOBOT with grant FP7-ICT-2011-9 2.1 – 600796.

References

1. Agris, U., Zieren, J., Canzler, U., Bauer, B., Kraiss, K.F.: Recent developments in visual sign language recognition. Univ. Access Inf. Soc. **6**, 323–362 (2008)
2. Antonakos, E., Pitsikalis, V., Maragos, P.: Classification of extreme facial events in sign language videos. EURASIP J. Image Video Process. **2014**, 14 (2014)
3. Arbib, M.A.: How the Brain Got Language: The Mirror System Hypothesis. Oxford University Press, New York (2012)
4. Bayer, I., Silbermann, T.: A multi modal approach to gesture recognition from audio and video data. In: Proceedings of the ACM International Conference on Multimodal Interaction, Sydney, pp. 461–466 (2013)
5. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, New York (2006)
6. Bolt, R.A.: Put-that-there: voice and gesture at the graphics interface. ACM Comput. Graph. **14**(3), 262–270 (1980)

7. Bordier, C., Puja, F., Macaluso, E.: Sensory processing during viewing of cinematographic material: computational modeling and functional neuroimaging. NeuroImage **67**, 213–226 (2013)
8. Bowden, R., Windridge, D., Kadir, T., Zisserman, A., Brady, M.: A linguistic feature vector for the visual interpretation of sign language. In: Proceedings of the European Conference on Computer Vision (ECCV), Prague (2004)
9. Buehler, P., Everingham, M., Zisserman, A.: Learning sign language by watching TV (using weakly aligned subtitles). In: Proceedings of the IEEE International Conference on Computer Vision & Pattern Recognition (CVPR), Miami, pp. 2961–2968 (2009)
10. Chow, Y.-L., Schwartz, R.: The N-best algorithm: an efficient procedure for finding top N sentence hypotheses. In: HLT'89 Proceedings of the Workshop on Speech and Natural Language, Morristown, pp. 199–202 (1989)
11. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. IEEE Trans. Pattern Anal. Mach. Intell. **23**(6), 681–685 (2001)
12. Cour, T., Sapp, B., Nagle, A., Taskar, B.: Talking pictures: temporal grouping and dialog-supervised person recognition. In: Proceedings of the IEEE International Conference on Computer Vision & Pattern Recognition (CVPR), San Francisco (2010)
13. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of the IEEE International Conference on Computer Vision & Pattern Recognition (CVPR), San Diego, pp. 886–893 (2005)
14. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn. Wiley, New York (2001)
15. Emmorey, K.: Language, Cognition, and the Brain: Insights from Sign Language Research. Lawrence Erlbaum Associates, Mahwah (2002)
16. Escalera, S., Gonzàlez, J., Baró, X., Reyes, M., Guyon, I., Athitsos, V., Escalante, H., Sigal, L., Argyros, A., Sminchisescu, C., Bowden, R., Sclaroff, S.: ChaLearn multi-modal gesture recognition 2013: grand challenge and workshop summary. In: Proceedings of the ACM International Conference on Multimodal Interaction, Sydney, pp. 365–368 (2013)
17. Escalera, S., Gonzàlez, J., Baró, X., Reyes, M., Lopes, O., Guyon, I., Athistos, V., Escalante, H.J.: Multi-modal gesture recognition challenge 2013: dataset and results. In: Proceedings of the ACM International Conference on Multimodal Interaction, pp. 445–452 (2013)
18. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.A.: Describing objects by their attributes. In: Proceedings of the IEEE International Conference on Computer Vision & Pattern Recognition (CVPR), Miami (2009)
19. Fei-Fei, L., Perona, P.: A Bayesian hierarchical model for learning natural scene categories. In: Proceedings of the IEEE International Conference on Computer Vision & Pattern Recognition (CVPR), San Diego (2005)
20. Gersho, A., Gray, R.M.: Vector Quantization and Signal Compression. Springer Science & Business Media, Boston (1992)
21. Glotin, H., Vergyr, D., Neti, C., Potamianos, G., Luettin, J.: Weighting schemes for audio-visual fusion in speech recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Salt Lake City, pp. 173–176 (2001)
22. Jaimes, A., Sebe, N.: Multimodal human–computer interaction: a survey. Comput. Vis. Image Underst. **108**(1), 116–134 (2007)
23. Jelinek, F.: Statistical Methods for Speech Recognition. MIT Press, Cambridge (1997)
24. Johnson, R.E., Liddell, S.K.: A segmental framework for representing signs phonetically. Sign Lang. Stud. **11**(3), 408–463 (2011)
25. Kendon, A.: Gesture: Visible Action as Utterance. Cambridge University Press, Cambridge/New York (2004)
26. Kopp, S., Bergmann, K.: Automatic and strategic alignment of co-verbal gestures in dialogue. In: Wachsmuth, I., de Ruiter, J., Kopp, S., Jaecks, P. (eds.) Alignment in Communication: Towards a New Theory of Communication, pp. 87–107. John Benjamins Publ. Co., Amsterdam (2013)

27. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: Proceedings of the IEEE International Conference on Computer Vision & Pattern Recognition (CVPR), Anchorage (2008)
28. Liddell, S.K.: Grammar, Gesture and Meaning in American Sign Language. Cambridge University Press, Cambridge (2003)
29. Maragos, P., Gros, P., Katsamanis, A., Papandreou, G.: Cross-modal integration for performance improving in multimedia: a review. In: Maragos, P., Potamianos, A., Gros, P. (eds.) Multimodal Processing and Interaction: Audio, Video, Text, pp. 3–48. Springer, New York (2008)
30. McNeill, D.: Gesture: a psycholinguistic approach. In: The Encyclopedia of Language and Linguistics, pp. 1–15. Elsevier, Boston (2006)
31. McNeill, D.: Gesture-speech unity: phylogenesis, ontogenesis microgenesis. Lang. Interact. Acquis. **5**(2), 137–184 (2014)
32. Ong, S., Ranganath, S.: Automatic sign language analysis: a survey and the future beyond lexical meaning. IEEE Trans. Pattern Anal. Mach. Intell. **27**, 873–891 (2005)
33. Ostendorf, M., Kannan, A., Austin, S., Kimball, O., Schwartz, R., Rohlicek, J.R.: Integration of diverse recognition methodologies through reevaluation of N-best sentence hypotheses. In: HLT'91 Proceedings of the Workshop on Speech and Natural Language, pp. 83–87 (1991)
34. Oviatt, S., Cohen, P.: Perceptual user interfaces: multimodal interfaces that process what comes naturally. Commun. ACM **43**(3), 45–53 (2000)
35. Parikh, D., Grauman, K.: Relative attributes. In: Proceedings of the International Conference on Computer Vision (ICCV), Barcelona (2011)
36. Pastra, K.: COSMOROE: a cross-media relations framework for modelling multimedia dialectics. Multimed. Syst. **14**, 299–323 (2008)
37. Pavlakos, G., Theodorakis, S., Pitsikalis, V., Katsamanis, A., Maragos, P.: Kinect-based multimodal gesture recognition using a two-pass fusion scheme. In: Proceeding of the IEEE International Conference on Image Processing (ICIP), Paris, pp. 1495–1499 (2014)
38. Pitsikalis, V., Katsamanis, A., Theodorakis, S., Maragos, P.: Multimodal gesture recognition via multiple hypotheses rescoring. J. Mach. Learn. Res. **16**, 255–284 (2015)
39. Pitsikalis, V., Theodorakis, S., Vogler, C., Maragos, P.: Advances in phonetics-based sub-unit modeling for transcription alignment and sign language recognition. In: Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition Workshops, Colorado Springs (2011)
40. Rabiner, L.R., Juang, B.H.: Fundamentals of Speech Recognition. Prentice Hall, Englewood Cliffs (1993)
41. Rose, R.C., Paul, D.B.: A hidden Markov model based keyword recognition system. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Albuquerque, pp. 129–132 (1990)
42. Searle, J.R.: Mind, Language, and Society: Philosophy in the Real World. Basic Books, New York (1999)
43. Sivic, J., Russell, B.C., Efros, A.A., Zisserman, A., Freeman, W.T.: Discovering objects and their location in images. In: Proceedings of the International Conference on Computer Vision (ICCV), Beijing, (2005)
44. Starner, T., Weaver, J., Pentland, A.: Real-time American sign language recognition using desk and wearable computer based video. IEEE Trans. Pattern Anal. Mach. Intell. **20**(12), 1371–1375 (1998)
45. Theodorakis, S., Pitsikalis, V., Maragos, P.: Dynamic–static unsupervised sequentiality, statistical subunits and lexicon for sign language recognition. Image Vis. Comput. **32**, 533–549 (2014)
46. Theodoridis, S., Koutroumbas, K.: Pattern Recognition, 4th edn. Academic Press (2008)
47. Tomasello, M.: Origins of Human Communication. MIT Press, Cambridge (2008)
48. Vatakis, A., Spence, C.: Audiovisual synchrony perception for music, speech, and object actions. Brain Res. **1111**, 134–142 (2006)

49. Vogler, C., Metaxas, D.: A framework for recognizing the simultaneous aspects of American sign language. Comput. Vis. Image Underst. **81**(3), 358–384 (2001)
50. Wilpon, J., Rabiner, L.R., Lee, C.H., Goldman, E.R.: Automatic recognition of keywords in unconstrained speech using hidden Markov models. IEEE Trans. Acoust. Speech Signal Process. **38**(11), 1870–1878 (1990)
51. Wittgenstein, L.: Philosophical Investigations. (Translated by Anscombe, G.E.M., and Editors Hacker, P.M.S., Schulte, J., 4th edn.). Wiley-Blackwell Publ. (2009) (1953)
52. Wittgenstein, L.: The Big Typescript: TS 213 (Edited and translated by Luckhardt, C.G., Aue, M.E.). Blackwell Publication (2005)
53. Wu, J., Cheng, J., Zhao, C., Lu, H.: Fusing multi-modal features for gesture recognition. In: Proceedings of the ACM International Conference on Multimodal Interaction, Sydney, pp. 453–460 (2013)