Gestural Interfaces for Hearing-Impaired Communication

Oya Aran, Thomas Burger, Lale Akarun, Alice Caplier

Gestural Interfaces for Hearing-Impaired Communication

Oya Aran[1], Thomas Burger[2], Lale Akarun[1], Alice Caplier[2]

[1]Dep. of Computer Engineering, Bogazici University 34342 Istanbul, Turkey, aranoya@boun.edu.tr, akarun@boun.edu.tr

[2]GIPSA-lab/DIS, 46 avenue Félix Viallet, 38031 Grenoble cedex 1, France, thomas.burger@lis.inpg.fr, alice.caplier@lis.inpg.fr

Abstract. Gestural interfaces, besides providing natural means of human-computer interaction for everyone, enable the hearing impaired to use sign language or better understand speech through vision. This chapter overviews (1) the various modalities involved in gestured languages (2) the mean to automatically apprehend them individually and (3) to fuse them in order to provide a communication medium adapted to hearing-impaired. We present two example applications, a sign language tutoring tool and a cued speech interpreter and discuss theoretical and practical aspects.

Keywords. Hand gesture recognition, belief functions, multimodal fusion, sign language, cued speech

Introduction

Recent research in Human-Computer Interaction (HCI) has focused on equipping machines with means of communication that are used between humans, such as speech and accompanying gestures. For the hearing im-

paired, the visual components of speech, such as lip movements, or gestural languages such as sign language are available means of communication. This has led researchers to focus on lip reading, sign language recognition, finger spelling, and synthesis. Gestural interfaces for translating sign languages, cued speech translators, finger spelling interfaces, gesture controlled applications, and tools for learning sign language have been developed in this area of HCI for the hearing impaired.

Gestural interfaces developed for hearing impaired communication are naturally multimodal. Instead of using audio and visual signals, hearing impaired people use multiple vision based modalities such as hand movements, shapes, position, head movements, facial movements and expressions, and body movements in parallel to convey their message.

(a)

(b)

(c)

Fig. 1. (a) An example sign “anne (mother)” from Turkish Sign Language (TSL), (b) Fingerspelling alphabet of TSL (Dynamics are indicated by red arrows), and (c) an example of French cued speech “bonjour (good morning)".

The primary means of communication between hearing-impaired people are sign languages. Almost all countries and sometimes regions within countries have unique sign languages that are not necessarily related with the spoken language of the region. Each sign language has its own grammar and rules (Stokoe 1960). Instead of audio signals, sign languages use hand movements, shapes, orientation, position, head movements, facial expressions, and body movements both in sequential and parallel order (Lidell 2003). Research on automatic sign language communication has progressed in recent years. Several survey papers are published that show the significant progress in the field (Ong and Ranganath 2005; Parton 2006). Interfaces are developed that handle isolated (Keskin et al. 2007) and continuous sign language recognition (Fang et al. 2007; Holden et al. 2005). Interactive educational tools have also been developed for teaching sign language (Aran et al. 2006).

Fingerspelling is a way to code the words with a manual alphabet which is a system of representing all the letters of an alphabet, using only the hands. Fingerspelling is a part of sign languages and is used for different purposes. It may be used to represent words which have no sign equivalent, or for emphasis, clarification, or when teaching or learning a sign language (Feris et al. 2004, Wu and Gao 2001).

Cued Speech (CS) is a more recent and totally different means of communication, whose importance is growing in the hearing-impaired community. It was developed by Dr. Cornett in 1967 (Cornett 1967). Its purpose is to make the natural oral language accessible to hearing-impaired people, by the extensive use of lip-reading. But lip-reading is ambiguous: for example, /p/ and /b/ are different phonemes with identical lip shape. Cornett suggests replacing invisible articulators (such as vocal cords) that participate to the production of the sound by hand gestures. Basically, it means completing the lip-reading with various manual gestures. Then, considering both lip shapes and hand gestures, each phoneme has a specific visual aspect. There are three modalities in CS: lip motion, hand shape and hand location with respect to the face.

Fig. 2 shows the overall architecture of the multimodal gesture based interfaces for the hearing impaired communication. In the next section, we discuss and review analysis techniques for the modalities that are used in hearing impaired communication. We concentrate on the individual modalities: hand, face, lips, expression and treat their detection, segmentation, and feature extraction. In the *Temporal analysis* section, we focus on the temporal analysis of the modalities, specifically in sign languages and in CS. The following section presents temporal modeling and belief-based multimodal fusion techniques. In the last section, we give two example applications: a sign language tutoring tool and a cued speech interpreter.

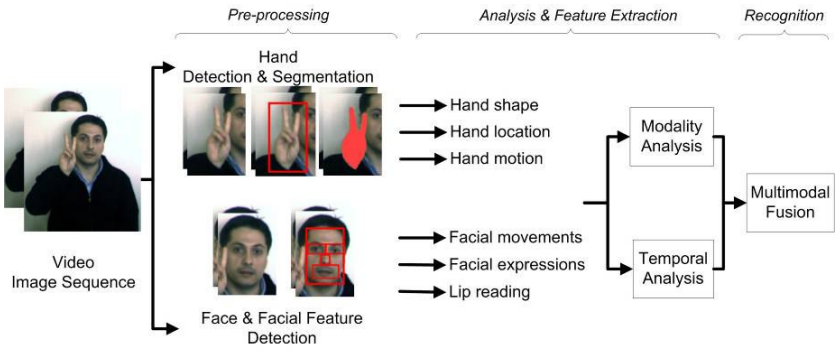


Fig. 2. Multimodal gestural interfaces for the hearing impaired

Modality Processing and Analysis

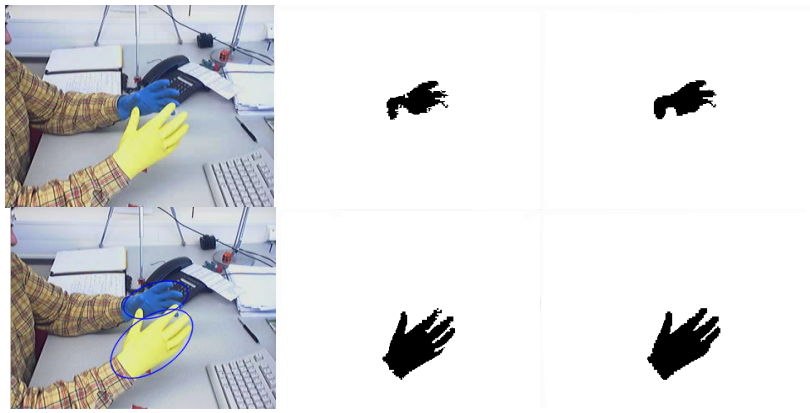
The modalities involved in gestured languages can be discussed from several points of view:

- *The part of the body that is involved*: Hands, head, facial features, shoulders, general standing, etc. For example, sign languages use the whole upper body, hands, head, facial features, and body/shoulder motion, whereas in cued speech, only a single hand and lips are in action.
- *Whether the modality conveys the main message or a paralinguistic message*: The hand shapes, locations and the lips in a CS phrase jointly convey the main message. On the other hand, in sign languages, paralinguistic elements can be added to the phrase via the non-manual elements or the variations of the manual elements. In sign languages, the main message is contained jointly in the manual (hand motion, shape, orientation and position) and non-manual (facial features, head and body motion) modalities where the non-manual elements are mainly used to complement, support or negate the manual meaning.
- *Whether the modality has a meaning by itself or not*: In CS, both modalities contain an ambiguity if they are used independently. The hand shapes code the consonants and the hand locations code the vowels. A hand shape-location pair codes several phonemes that are differentiated by the lip shape. In sign languages, the manual component has a meaning by itself for most of the signs. For a small number of signs, the non-manual component is needed for full comprehension.

In this section, we present analysis and classification methods for each of the modalities independently. The synchronization, correlation, and the fusion of modalities are discussed in the next sections.

Preprocessing

Vision based systems for gestural interfaces provide a natural environment in contrast to the cumbersome instrumented gloves with several sensors and trackers that provide accurate data for motion capture. However, vision based capture methodology introduces its own challenges, such as the accurate detection and segmentation of the face and body parts, hand and finger configuration, or handling occlusion. Many of these challenges can be overcome by restricting the environment and clothing or by using several markers such as differently colored gloves on each hand or colored markers on each finger and body part. In communication of the hearing impaired, the main part of the message is conveyed through the hands and the face. Thus the detection and segmentation of hands and face in a vision based system is a very important issue.



(a) (b) (c)

Fig. 3. Hand segmentation by automatically defined thresholds. (a) Original image and the detected hands, (b) thresholding & connected components labeling, (c) region growing

Hand detection

Hand detection and segmentation can be done with or without markers. Several markers are used in the literature such as single colored gloves on

each hand, or gloves with different colors on each finger or joint. With or without a marker, descriptors of color, motion and shape information, separately or together, can be used to detect hands in images (Habili et al. 2004; Holden et al. 2005; Awad et al. 2006). Similar techniques are used to detect skin colored pixels or the pixels of the glove color. Color classification can be done either parametrically or non-parametrically. In parametric methods, a distribution is fitted for the color of interest, and the biggest connected region is extracted from the image (see Fig. 3 from Aran and Akarun 2006).

A popular non-parametric method is histogram-based modeling of the color (Jayaram et al. 2004). In this approach, a normalized histogram is formed using the training pixels and the thresholds are determined. The similarity color map of the image is extracted using the histogram bins. Similar steps, thresholding, connected components labeling and region growing, are applied to obtain the segmented hand (Aran et al. 2006).

The main advantage of using a marker is that it makes tracking easier and helps to resolve occlusions. In a markerless environment, hand tracking presents a bigger challenge. In sign languages, the signing takes place around the upper body and very frequently near or in front of the face. Moreover the two hands are frequently in contact and often occlude each other. Another problem is to decide which of these two hand regions correspond to the right and left hands. Similarly in CS, the frequent face/hand contacts are difficult to deal with. Thus, the tracking and segmentation algorithm should be accurate enough to resolve these conditions and provide the two segmented hands.

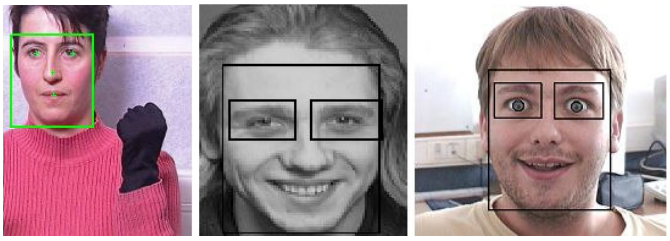


Fig. 4. Face localization

Face detection

The user face conveys both verbal and non-verbal communication information. The first step is to localize the user's face during the gesture analysis process. Face localization have been widely studied (Hjelmäs and Low 2001, Yang et al. 2002). The most popular face detector is the detector developed by (Viola and Jones 2004) whose code is freely available

(MPT). Independent of the technique employed, the output of the face detector is a bounding box around the face and the position of some facial features, as shown in Fig. 4.

Retinal pre-processing

In the human retina, some low level processing is done on video data. This processing is very efficient in order to condition the data for high level processing.

In the human retina (Bullier 2001), two steps of filtering (OPL and IPL filtering) are done so that two information channels are extracted: the Parvo (parvocellular) channel dedicated to detail analysis (details: static contours enhancement) and the Magno (magnocellular) channel dedicated to motion analysis (moving contours enhancement). For a more detailed description of the retina modeling and properties, see (Benoit and Caplier 2005a, 2005b, Herault 2007).

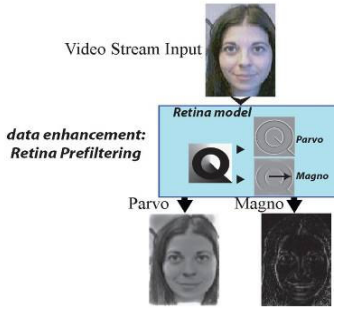


Fig. 5. Retina filtering effect

In the sequel of this chapter, we provide several examples where the properties of the retina are used to condition video data before high level processing.

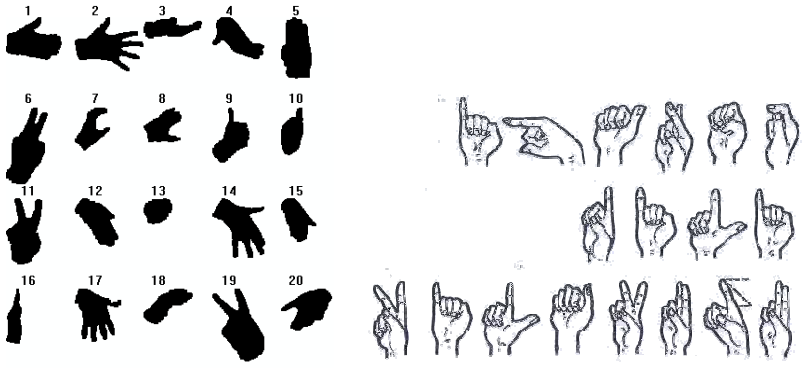
Hand shape

Hand shape is one of the main modalities of the gestured languages. Apart from sign language, CS or finger spelling, hand shape modality is widely used in gesture controlled computer systems where predefined hand shapes are used to give specific commands to the operating system or a program. Analysis of the hand shape is a very challenging task as a result of the very high degree of freedom of the hand. For systems that use limited number of simple hand shapes, such as hand gesture controlled systems (hand

shapes are determined manually by the system designer) or in CS (the French, Spanish, English and American CSs are based on eight predefined hand shapes), the problem is easier. However, for sign languages, the unlimited number and the complexity of the hand shapes make discrimination a very challenging task, especially with 2D vision based capture systems.

Fig. 6. French cued speech hand shapes

In CSs, there are eight hand shapes that mainly differ by open and closed fingers. The CSs coding is ideally performed in 2D. Thus, the hand is supposed to be frontal, and all the hand rotations are supposed to be planar, although it is not the case in practical situations (which is the source of one of the main difficulties). French cued speech (FCS) hand shapes are presented in Fig. 6.



(a) (b)

Fig. 7. Example hand shapes from (a) ASL and (b) TSL

In sign languages, the number of hand shapes is much higher. For example, without considering fingerspelling, American Sign Language (ASL) has around 150 hand shapes, and in British Sign Language there are 57 hand shapes. These hand shapes can be further grouped into around 20 phonemically distinct subgroups. Example hand shapes from ASL and TSL are given in Fig. 7.

Inertial Study of the Hand

It is possible to compute the global direction of the hand shape using principal axis analysis. Then, a hand rotation in order to work on a vertical shape is considered in order to make the whole study easier.

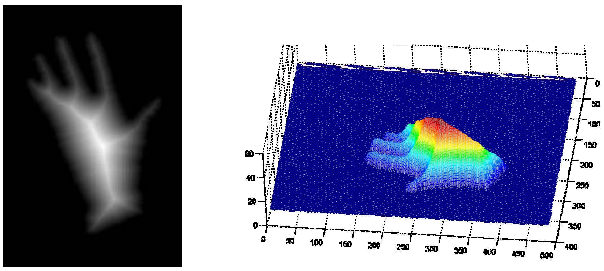
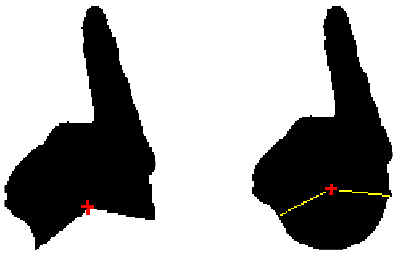


Fig. 8. Illustration of the distance transform in gray level (black pixels belong to the background and the lighter the pixels the higher its value after the distance transform) and in 3D.

The Distance Transform of the binary image of an object associates to each pixel of the object its distance to the closest pixel of the background, and associates the value 0 to all the pixels of the background.

Obviously, the centre of palm is one of the points of the hand which is the furthest from the contour of the hand (Fig. 8). As a consequence, the palm of the hand can be approximated by a circle whose radius is related to the maximum value given by the distance transform of the binary hand image.



(a) (b)

Fig. 9. (a) Palm delimitation, (b) once the "V" is removed, the shape is instable, (c) after the "V" is filled with a disc whose radius linearly varies between the two sides of the "V".

Once the palm is approximated by a circle, the wrist (or eventually the forearm) is removed as illustrated in Fig. 9.

The next step is to detect particular fingers. The main application is the study of a pointing (or deixis) gesture. The deixis gesture may be performed by the whole arm, and sometimes by the gaze of the eye. However, we consider hand shape for the pointing gestures:

- The general hand shape orientation is used to indicate a direction. In such a case, it is straightforward to deal with as the first principal axis of the bounding box corresponds to the deixis direction.
- The longest unfolded finger is used to materialize a pointing zone (for instance, a cursor gesture for HMI).
- The position of the extremity of a particular finger is considered depending on the hand shape.
- The precise deixis gesture with a single finger is replaced by a deixis gesture where the pointing element does not belong to the hand, but to its convex hull (linear or polynomial). This case is practically very likely in human gestures, including CSs, for which the pointing rules are supposed to be really strict.

In the case of CSs, the deixis gesture is of prime importance, as it is required to determine the location of the hand with respect to the face.

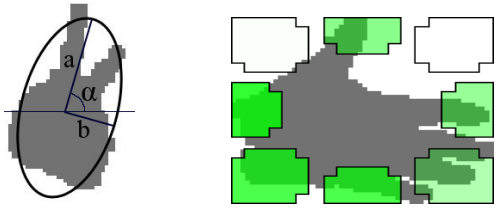
The location is determined by the position of the pointing finger with respect to the face. It is theoretically the longest one, but, in practice, (1) parallax errors, (2) wrist flexion, and (3) the use of the convex hull of the hand shape modify the problem. Then a more robust algorithm, using fusion of information from hand shape and respective positions, must be used (Burger 2007).

In (Burger et al. 2007b), we consider the use of a thumb presence indicator, which returns a non-zero positive value if the thumb is unfolded and 0 otherwise. This is useful when (1) the thumb-up gesture is used, or when (2) the thumb presence has a particular meaning. The approach uses the polar parametric representation of the binary hand shape. The peaks of this representation correspond to potential fingers of the hand shape. Thresholds, derived from statistics on the morphology of the hand (Norkin and Levangie 1992), are defined in order to materialize the region of the thumb extremity when it is unfolded. If a finger is detected within the thumb area, then, it is the thumb. The corresponding peak height is measured with respect to the lowest point between the thumb and the index. This value provides a numerical indicator of the presence of the thumb.

Hand shape descriptors

To analyze the hand shape, appearance or 3D-model based features can be used (Wu and Huang 2001). Appearance based features are preferred due to their simplicity and low computation times, especially for real time ap-

plications. Region based descriptors (image moments, image eigenvectors, Zernike moments, Hu invariants, or grid descriptors) and edge based descriptors (contour representations, Fourier descriptors, or Curvature Scale Space descriptors) can be used for this purpose. A survey on binary image descriptors can be found in (Zhang 2003).



(a) (b)

Fig. 10. (a) Best fitting ellipse, (b) Area filters. Green and white pixels indicate the areas with and without and hand pixels, respectively.

A combination of several appearance based features is used as hand shape features for recognizing ASL signs (Aran et al. 2006). Half of the features are based on the best fitting ellipse (in least-squares sense) to a binary image, as seen in Fig. 10a. The rest are calculated using area filters as seen in Fig. 10b. The bounding box of the hand is divided into eight areas, in which percentage of on and off hand pixels are calculated.

Hu invariants are successful in representing hand shapes (Hu 1962; Caplier et al. 2004; Burger et al. 2007b). Their purpose is to describe the binary shape region via several moments of various orders, on which specific transforms ensure invariance properties.

The centered scale invariant inertial moments of order $p+q$ are calculated as follows:

$$n_{pq} = \frac{m_{pq}}{m_{00}^{\frac{p+q}{2}+1}} \quad \text{with} \quad m_{pq} = \iint_{x\,y} (x-\bar{x})^p \, (y-\bar{y})^q \, \delta(x,y)\, dx\, dy$$

where $\bar{x}$ and $\bar{y}$ are the coordinates of the center of gravity of the shape and $\delta(x,y)=1$ if the pixel belongs to the hand shape and 0 otherwise. Then, the seven Hu invariants are calculated:

$$S_1 = n_{20} + n_{02}$$
$$S_2 = (n_{20} + n_{02})^2 + 4 \cdot {n_{11}}^2$$
$$S_3 = (n_{30} - 3 \cdot n_{12})^2 + (n_{03} - 3 \cdot n_{21})^2$$
$$S_4 = (n_{30} + n_{12})^2 + (n_{03} + n_{21})^2$$
$$S_5 = (n_{30} - 3 \cdot n_{12}) \cdot (n_{30} + n_{12}) \cdot \left((n_{30} + n_{12})^2 - 3 \cdot (n_{03} + n_{21})^2\right) - (n_{03} - 3 \cdot n_{21}) \cdot (n_{03} + n_{21}) \cdot \left(3 \cdot (n_{30} + n_{12})^2 - (n_{03} + n_{21})^2\right)$$
$$S_6 = (n_{20} + n_{02}) \cdot \left((n_{30} + n_{12})^2 - (n_{03} + n_{21})^2\right) + 4 \cdot {n_{11}}^2 \cdot (n_{30} + n_{12}) \cdot (n_{03} + n_{21})$$
$$S_7 = (3 \cdot n_{21} - n_{03}) \cdot (n_{30} + n_{12}) \cdot \left((n_{30} + n_{12})^2 - 3 \cdot (n_{03} + n_{21})^2\right) - (n_{30} - 3 \cdot n_{12}) \cdot (n_{03} + n_{21}) \cdot \left(3 \cdot (n_{30} + n_{12})^2 + (n_{03} + n_{21})^2\right)$$

We have used Hu invariants as descriptors of CS hand shapes (Burger et al. 2007b). The experiments show that Hu invariants have an acceptable performance which can be improved by the addition of the thumb information presence.

The Fourier-Mellin Descriptors (FMD) are an interesting alternative (Adam et al. 2001) . The Fourier-Mellin Transform (FMT) of a function f corresponds to its Mellin transform result represented in terms of Fourier coefficients. The FMT is defined for all real positive function $f(r,\theta)$ in polar coordinates (the shape to describe) so that the Mellin transform is 2π-periodic:

$$M_f(q,s) = \int_{r=0}^{\infty} \int_{\theta=0}^{2\pi} r^{s-1} e^{-iq\theta} f(r,\theta)\, dr d\theta \quad \text{with} \quad q \in \mathbb{Z},\ s = \sigma + i\upsilon \in \mathbb{C},\ and\ i = \sqrt{-1}$$

Then the application of the delay theorem and the extraction of the module of the FMT lead to a set of descriptors indexed by q and s. They are rotation invariant, and normalization by $M_f(0, \sigma)$ makes them scale invariant. The translation invariance is derived from the choice of the centre of development (the origin of (r,θ) coordinates).

In case of digital images, it is necessary to digitalize the FMT and to convert the sampled Cartesian space into a polar space. In practice, $M_f(0, \sigma)$ is approximated by:

$$M_f(q,\sigma + i\upsilon) \approx \sum_{\substack{k,l \\ 0 \le (k^2+l^2) \le r^2_{\max}}} h_{p,q}(k,l) \cdot f\left(\bar{k} - k, \bar{l} - l\right)$$

with
- $\left(\bar{k}, \bar{l}\right)$ centre of development of *the* FMT (here, the gravity centre of the hand)
- $r_{\max}$ superior bound for r
- $h_{p,q}(k,l) = \dfrac{1}{(k^2+l^2)^{1-\frac{\sigma}{2}}} \cdot \exp\left(i.\left(\dfrac{p}{2}\ln(k^2+l^2) - q.\arctan\left(\dfrac{l}{k}\right)\right)\right)$

These descriptors are particularly efficient to discriminate hand shapes, even in cases of (1) multi-coder (when the morphologic variability is introduced), (2) unknown coder, (3) imprecise classifier tuning (Burger 2007).

Hand location

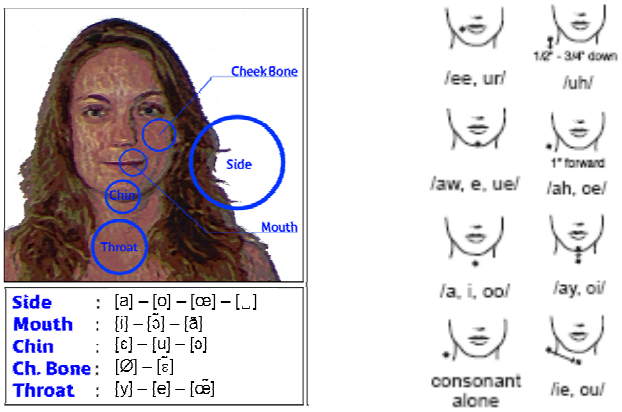
The location of the hand must be analyzed with respect to the context. It is important to determine the reference point on the space and on the hand. In sign languages, where both the relative location and the global motion of the hand are important (see Fig. 11), the continuous coordinates and the location of the hand with respect to body parts should be analyzed. This analysis can be done by using the center of mass of the hand. On the other hand, for pointing signs, using center of mass is not appropriate and the coordinates of the pointing finger and the pointing direction should be used. Since signs are generally performed in 3D space, location analysis should be done in 3D if possible. Stereo cameras can be used to reconstruct 3D coordinates in vision based systems.

CLEAN sign: Hand location w.r.t the other hand

DRINK sign: Hand location w.r.t the mouth

Fig. 11. Possible reference points on the signing space. Examples from ASL.

The hand locations in CS are determined by the pointing location of the hand with respect to the coder's face. For example, in French CS, "mouth", "side", "throat", "chin", and "cheek bone" are used as five different locations on the face (see Fig. 12a). Once the pointing finger and the face features are located of the image, determining the location component of the gesture is rather simple.



(a) (b)

Fig. 12. (a) French and (b) American cued speech hand locations, and their phonemic meaning.

Hand motion

In gestured communication, it is important to determine whether the performed hand motion conveys a meaning by itself.

From a linguistic point of view, FCS is the complete visual counterpart of oral French. Hence, it has a comparable prosody and the same dynamic aspect. From a gesture recognition point of view, the interpretation is completely different: each FCS gesture {Hand shape + Location} is a static gesture (named a *target gesture*) as it does not contain any motion and can be represented in a single picture or a drawing. Then, a coder is supposed to perform a succession of targets. In real coding, the hand nevertheless moves from target to target (as the hand cannot simply appear and disappear) and *transition gestures* are produced. We consider that FCS is inherently static: target images are sufficient to decode the continuous sentence: as a consequence, complete transition analysis is most of the time useless to be processed (Burger 2007; Burger et al. 2007a).

In sign languages, the hand motion, together with the hand shape and location, is one of the primary modalities that form the sign. Depending on the sign, the characteristic of the hand trajectory can change, requiring different levels of analysis. For example, some signs are purely static and there is no need for trajectory analysis. The motion of the dynamic signs can be examined as either of two types:

1. Signs with global hand motion: In these signs, the hand center of mass translates in the signing space.
2. Signs with local hand motion: This includes signs where the hand rotates without any translation, or where the finger configuration of the hand changes.

Trajectory Analysis

Trajectory analysis is needed for signs with global hand motion. For signs with local motion, the change of the hand shape over time should be analyzed in detail, since even small changes of the hand shape convey information content.

The first step of hand trajectory analysis is tracking the center of mass of each segmented hand. Hand trajectories are generally noisy due to segmentation errors resulting from bad illumination or occlusion. Thus a filtering and tracking algorithm is needed to smooth the trajectories and to estimate the hand location when necessary. Moreover, since hand detection is a costly operation, hand detection and segmentation can be applied not in every frame but less frequently, provided that a reliable estimation algorithm exists. For this purpose, algorithms such as Kalman filters and particle filters can be used. Kalman filters are linear systems with Gaussian noise assumption and the motion of each hand is approximated by a constant velocity or a constant acceleration motion model. Particle filtering, also known as the condensation algorithm (Isard and Blake 1998), is an alternative with non-linear and non-Gaussian assumptions. The main disadvantage is its computational cost which prevents its usage in real time systems.

Based on the context and the sign, hand coordinates should be normalized with respect to the reference point of the sign, as discussed in the previous section. In addition to the coordinates, the velocity and the acceleration can be used as hand motion features.

Several methods are proposed and applied in the literature for modeling the dynamics of signs or hand gestures. These include Hidden Markov Models (HMM) and its variants, Dynamic Time Warping (DTW), Time Delay Neural Networks (TDNN), Finite State Machines (FSM), and temporal templates. Some of these techniques are only suitable for simple hand gestures and cannot be applied to complex dynamic systems. Among dynamic systems, HMM and its variants are popular in sign language recognition, and hand gesture recognition in general.

Static Gestures in Dynamic Context

In order to take advantage of the static nature of some gestures, let us assume that it is possible to extract target gestures from the surrounding transition motions using low-level kinetic information that can be extracted before the complete recognition process.

This hypothesis is motivated by the analysis of FCS sequences, and can be generalized directly to other static gestural languages. It shows that the hand is slowing down each time the hand is reaching a phonemic target. As a consequence, target gestures have slower hand motion than transition gestures. It nonetheless appears that there is almost always some residual motion during the realization of the target (because of the co-articulation).

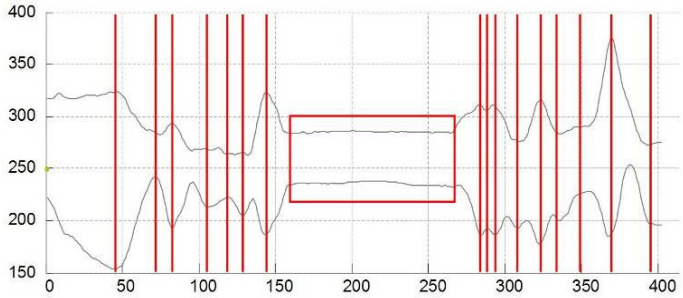


Fig. 13. Representation of the coordinates (vertical above and horizontal below) of the gravity centre of the hand shape during a CS sequence. The vertical lines correspond to images of minimal motion that are target images of hand location.

In case the motion in which the static gesture is hidden is a global translation motion (i.e. the change of location in CS or any deixis gesture), any study of the rigid movement is likely to stress the variations of speed and the images on which the motion is small enough to be potentially considered as a target gesture. Fig. 13 illustrates the trajectory of the hand gravity centre during a CS video sequence. It appears that each image for which the two components of the trajectory are stable (which corresponds to local minima in the speed evolution) corresponds to some location being reached.

In case of non-rigid motion, such as the deformation of the hand contour when its shape is considered, it is more difficult to define a cinematic clue that indicates when a target is reached or when the image represents a transitive motion. In order to do so, an approach based on the properties of the retina (and specially the IPL filter) has been proposed in (Burger et al. 2006a). A dedicated retina filter (Burger et al. 2007a) has been defined to evaluate the amount of deformation of the hand contour along the se-

quence. It is made of several elements which are chained together (Fig. 14). As established in (Burger 2007; Burger et al. 2007a) this method is particularly efficient.

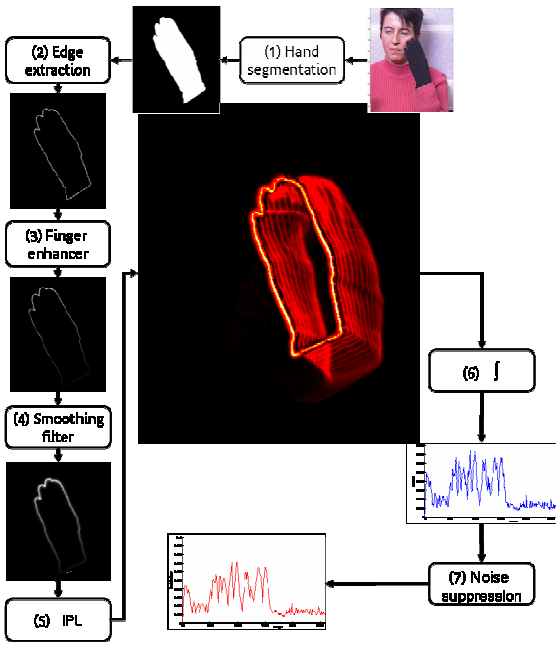


Fig. 14. Flowchart of the dedicated retina filter

Facial movements

Thanks to the retina model, it is possible to efficiently detect some facial movements. The analysis of the temporal evolution of the energy of the Magno output related to moving contours has been used in (Benoit and Caplier 2005c) in order to develop a motion events detector. Indeed, in case of motion, the Magno output energy is high and on the contrary, if no motion occurs, the Magno output energy is minimum or even null. In Fig. 15, the case of an eye blink sequence is illustrated: the motion event detector generates a signal $\alpha(t)$ which reaches 1 each time a blink is detected (high level of energy on the Magno channel, frames 27, 59 and 115) and which is 0 if no blinks are present (the energy of the Magno channel is null).

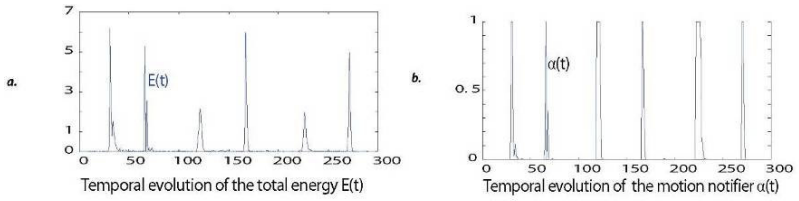


Fig. 15. a. Temporal evolution of the Magno output in case of a blink video sequence; b. temporal evolution of the motion events detector

Lip reading

The main difference between SL and CS is that the CS message is partly based on lip reading: Although signers also use lip movements while they are signing, it is not a part of the sign language. However, for CS, it is as difficult to read on the lip without any CS hand gesture, than to understand the hand gestures without any vision of the mouth. The link between lip and oral message is included in the shape and the motion of the lips.
An important step for lip reading is lip contours extraction. Significant research has been carried out to accurately obtain the outer lip contour. One of the most popular approaches is using snakes (Kass et al. 1988), which have the ability to take smoothing and elasticity constraints into account (Terzopoulos and Waters 1993; Aleksic et al. 2002). Another popular approach is using active shape models and appearance shape models. (Cootes 1994) presents statistical active model for both shape (AMS) and appearance (AAM). Shape and grey-level appearance of an object are learned from a training set of annotated images. Then, a Principal Component Analysis (PCA) is performed to obtain the main modes of variation. Models are iteratively matched to reduce the difference between the model and the real contour by using a cost function. Another approach is presented in (Eveno et al. 2004), where a parametric model associated with a "jumping snake" for the initialization phase is proposed.

Relatively few studies deal with the problem of inner lip segmentation. The main reason is that inner contour extraction from front views of the lips without any artifice is much more difficult than outer contour extraction. Indeed, we can find different mouth shapes and non-linear appearance variations during a conversation. Especially, inside the mouth, there are different areas which have similar color, texture or luminance than lips (gums and tongue). We can see very bright zones (teeth) as well as very dark zones (oral cavity). Every area could continuously appear and disappear when people are talking. Among the few existing approaches for inner lip contour extraction, lip shape is represented by a parametric deform-

able model composed of a set of curves. In (Zhang 1997), authors use deformable templates for outer and inner lip segmentation. The chosen templates are three or four parabolas, depending on whether the mouth is closed or open. The first step is the estimation of candidates for the parabolas by analyzing luminance information. Next, the right model is chosen according to the number of candidates. Finally, luminance and color information is used to match the template. This method gives results, which are not accurate enough for lip reading applications, due to the simplicity and the assumed symmetry of the model. In (Beaumesnil et al. 2006), authors use internal and external active contours for lip segmentation as a first step. The second step recovers a 3D-face model in order to extract more precise parameters to adjust the first step. A k-means classification algorithm based on a non-linear hue gives three classes: lip, face and background. From this classification, a mouth boundary box is extracted and the points of the external active contour are initialized on two cubic curves computed from the box. The forces used for external snake convergence are, in particular, a combination of non-linear hue and luminance information. Next, an inner snake is initialized on the outer contour, and then shrunk by a non isotropic scaling with regard to the mouth center and taking into account the actual thickness of the lips. The main problem is that the snake has to be initialized close to the contour because it will converge to the closest gradient minimum. Particularly for the inner lip contour, different gradient minima are generated by the presence of teeth or tongue and can cause a bad convergence. The 3D-face model is used to correct this problem, but the clone does not give accurate results for lip reading.

In (Luettin et al. 1996), an AMS is build and in (Gacon et al. 2005), an AMS and an AAM are built to inner and outer lip detection. The main interest of these models is that the segmentation gives realistic results, but the training data have to deal with many cases of possible mouth shapes.

Once the mouth contours have been extracted, lip shape parameters for lip reading have to be extracted. Front views of the lips are phonetically characterized with lip width, lip aperture and lip area. These lip parameters are derived from the inner and outer contours. In an automatic recognition task of lip-reading process, it is thus pertinent to consider these parameters

Facial expressions

A summary of the significant amount of research carried out in facial expression classification can be found in (Pantic et al. 2000) and (Fasel et al. 2003). One of the main approaches is optical flow analysis from facial actions (Yacoob and Davis 1996; Black and Yacoob 1997; Essa and Pentland

1997; Cohn et al. 1998]: These methods focus on the analysis of facial actions where optical flow is used to either model muscle activities or to estimate the displacements of feature points. A second approach is using model-based approaches (Zhang et al. 1998; Gao et al. 2003; Oliver et al. 2000; Abboud et al. 2004): Some of these methods apply an image warping process to map face images into a geometrical model. Others realize a local analysis where spatially localized kernels are employed to filter the extracted facial features. Once the model of each facial expression is defined, the classification consists in classifying the new expression to the nearest model using a suitable metric. A third group is fiducial points based approaches (Lien et al. 1998; Tian et al. 2001; Cohen et al. 2003; Tsapatsoulis et al. 2000): Recent years have seen the increasing use of geometrical features analysis to represent facial information. In these approaches, facial movements are quantified by measuring the geometrical displacement of facial feature points between the current frame and a reference frame.

We are going to illustrate the approach described in detail in (Hammal et al. 2007). In this work, the classification process is based on the Transferable Belief Model (TBM) (Smets and Kennes 1994) framework (see section on belief functions). Facial expressions are related to the six universal emotions, namely *Joy, Surprise, Disgust, Sadness, Anger, Fear,* as well as *Neutral*. The proposed classifier relies on data coming from a contour segmentation technique, which extracts an expression skeleton of facial features (mouth, eyes and eyebrows) and derives simple distance coefficients from every face image of a video sequence (see Fig. 16).

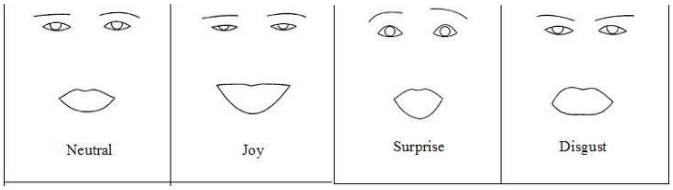


Fig. 16. Facial expression skeletons

The characteristic distances are fed to a rule-based decision system that relies on the TBM and data fusion in order to assign a facial expression to every face image. This rule-based method is well adapted to the problem of facial expression classification because it deals with confusing expressions (*Joy* or *Disgust*, *Surprise* or *Fear*, etc) and recognizes an *Unknown* expression instead of forcing the recognition of a wrong expression. Indeed, in the presence of doubt, it is sometimes preferable to consider that both expressions are possible rather than to choose one of them.

Temporal Analysis

In a multimodal interface, the correlation and synchronization of modalities must be clearly analyzed. This is a necessary step prior to multimodal fusion.

Sign Language

The temporal organization of sign languages can be analyzed in two: (1) The temporal organization within manual components (manual simultaneity), (2) the temporal organization between manual and non-manual components (manual/non-manual simultaneity).

The manual simultaneity is due to the usage of two independent moving elements: The two hands can perform different signs at the same time. We can classify the signs in a language as one or two-handed signs. In two-handed signs, the two hands are synchronized and perform a single sign. Whereas in one-handed signs, the performing hand is called the dominant hand and the other hand is idle in the isolated case. In continuous signing, as a result of the speed, while one hand is performing one sign, the other hand may perform the next sign, at the same time. From the recognition point of view, this property enforces the independent modeling of the two hands, while keeping their relation in case of two-handed signs.

The simultaneity of manual/non-manual components depends on the linguistic property of the performed sign. For example, non-manual signs for grammatical operators, such as negation and question, are performed over a phrase which generally includes more than one sign. On the other hand, the modifications on the meaning of a sign are performed via non-manual signs and they only affect the sign in focus. Of course, if these modifications affect a phrase, then the non-manual signs co-occur with one or more manual sign.

Cued Speech

In this section, we describe the temporal organization of the three modalities (hand shape, location, lips) of French Cued Speech. This description is based on the observation of numerous video sequences featuring a professional coder (hearing able translators) as well as hearing impaired people. A first study (Attina 2005) has been published by Attina on the desynchronization between the labial motion and the manual one, but the desynchronization of the two modalities of the manual motion (the hand shape

movement and the location movement) is not in its scope. Here, we summarize the principal results of (Attina 2005) and we complete them with observations about hand shape/ location temporal organization.

The main point of (Attina 2005) is a temporal scheme which synthesizes the structure of the code along time from a hand/lip delay point of view.

From this work it is possible to extract two remarks: The first is that the hand is in advance with respect to the lips, and apparently, the labial motion disambiguates the manual motion, and not the contrary. The second is that the variability of desynchronization is much too important to be directly used in a recognition system which automatically balances the desynchronization. Nevertheless, this scheme contains a lot of information which can be used to set the parameters of an inference system which purpose is to find a best matching between the modalities.

In general, the hand shape target is reached before the location target. This is easily explained by mechanic and morphologic arguments: in case of finger/face contact, the pointing finger must be completely unfolded before the beginning of the contact. As a consequence, hand shapes are in advance with respect to the locations. However, for some other hand shape/ location pairs, this observation is not valid (Burger 2007).As a consequence, it is really difficult to establish a precise enough model to forecast the desynchronization pattern. Nonetheless, the desynchronization are most of the time of intermediate amplitude (except at the beginning and the end of a sentence) so that computing a matching among the modalities in order minimize the desynchronization does not seem intractable.

Multimodal Fusion

There are two major difficulties in integrating modalities of gesture based communication: joint temporal modeling and multiplexing information of heterogeneous nature.

Temporal modeling

In gesture based communication of the hearing impaired, multiple modalities are used in parallel to form the linguistic units such as signs/words in sign languages or phonemes in CS. The temporal relation between these modalities must be carefully investigated to obtain a good temporal model that will result in high accuracies in a recognition task.

Hidden Markov Models

Among the temporal modeling techniques for hand gestures HMMs draw much attention (Rabiner 1998). Their success comes from their ability to cope with the temporal variability among different instances of the same sign.

HMMs are generative probability models that provide an efficient way of dealing with variable length sequences and missing data. Among different kinds of HMM architectures, left-to-right HMMs (Fig. 17) with either discrete or continuous observations are preferred for their simplicity and suitability to the hand gesture and sign language recognition problems.

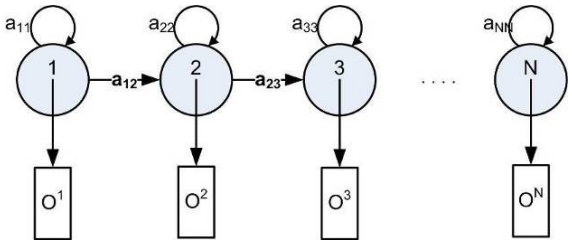


Fig. 17. Left-to-right HMM architecture

An HMM consists of a fixed number of states. Given a data sequence, the probabilities to determine the start state and transition probabilities, one can construct a state sequence. Each state generates an output (an observation) based on a probability distribution. This observation is the features observed at each frame of the data sequence.

For a sequence classification problem, one is interested in evaluating the probability of any given observation sequence, $\{O_1\ O_2\ ...\ O_T\}$, given a HMM model, Θ.

In isolated sign language recognition, an HMM model is trained for each sign in the dictionary. The simplest case is to put the features of all the concurrent modalities in a single feature vector. The likelihood of each model is calculated and the sequence is classified in to the class of the model that produces the highest likelihood. Instead of concatenating the features into a single feature vector, a process can be dedicated for each modality with established links between the states of different processes. In (Brand et al. 1997), Coupled HMMs are proposed for coupling and training HMMs that represent different processes (see Fig. 18a).

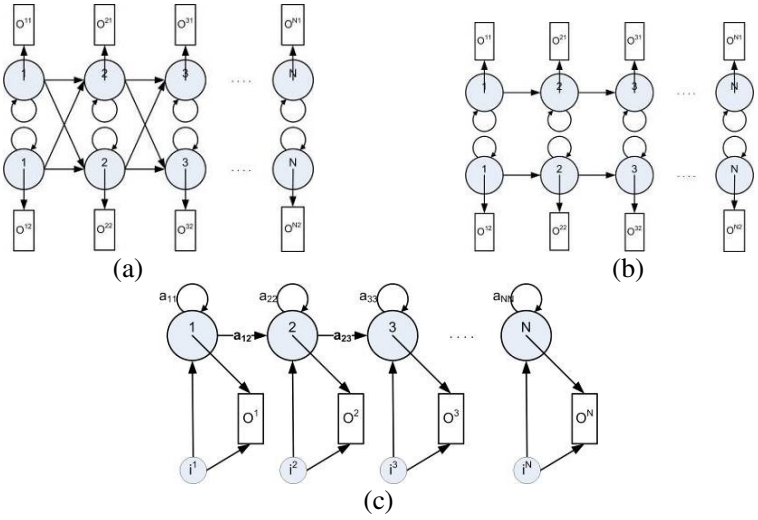


Fig. 18. (a) Coupled HMM, (b) Parallel HMM, (c) Input-Output HMM

When the synchronization of the modalities is weak, then it is not a good idea to process all the modalities in a single temporal model. Several models for each of the modalities can be used independently and integration can be done afterwards. An example is the Parallel HMM, as illustrated in Fig. 18b (Vogler and Metaxas 1999). Belief based methods can also be used to fuse different models to handle the ambiguity in between, as we describe in the following sections.

An alternative is to use Input Output HMMs (IOHMM) (see Fig. 18c) which model sequential interactions among processes via hidden input variables to the states of the HMM (Bengio and Frasconi 1996).

Co-articulation

In continuous gestural language, the consequent signs affect the beginning and end of each other. This co-articulation phenomenon can also be seen in spoken languages. When an HMM for each sign is trained to recognize the signs, the performance will drop down since each sign in the continuous signing will be slightly different than their isolated equivalents. Many of the methods proposed for solving the co-articulation affect, rely on modeling the co-articulation by using pairs or triples of signs instead of a single one.

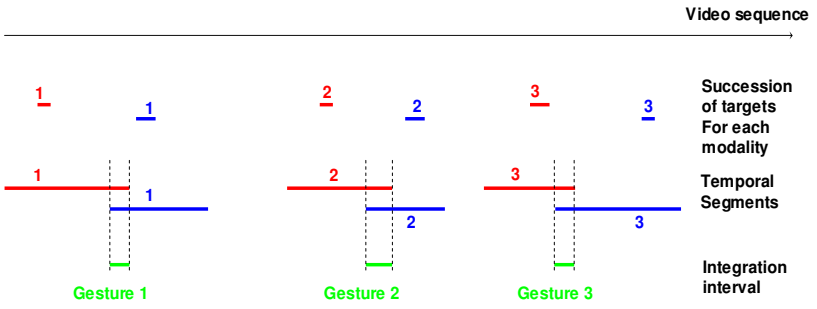


Fig. 19. Illustration of the definition of the temporal segments. Their overlapping deal with the dynamical aspect of the fusion of the modalities.

In case the modalities to be fused have a static nature which is classically hidden in a dynamic context because of a co-articulation phenomenon, we propose an alternative solution (Burger 2007). The main idea is to associate to any static gesture of the modalities a temporal segment of the video sequence which is centered on the target image. This segment is supposed to represent the time interval in which it is not possible to get another static gesture for a minimum time interval is necessary to produce the transition movements which are required to reach and to leave the target of the gesture. Then, whatever the recognition process output within this segment, it is reasonable to assume that a single gesture has been produced during this time interval. As a consequence, even if the target gestures of each modality are not produced at the same time, it is possible to balance this lack of synchronization by matching the segments which overlap (Fig. 19).

Of course, such a process only allows balancing small desynchronization. If the desynchronization is larger than the segments associated to the target images, it is impossible to easily warp the modalities. On the other hand, this hypothesis of small desynchronization is not that an important restriction. In HCI systems, it is rather common to assume that the "gesturer" (coder/signer) produces an academic motion, which means, he/she is concentrated on limiting the desynchronization between the various component of his/her gestures.

In the general case, if multiple overlaps and/or empty intersection remain too numerous to allow a simple matching, then, the use of DTW methods or graph matching algorithm can be successfully applied to finalize the temporal matching of the modalities.

Heterogenic Multiplexing

The purpose of fusing the various gestural modalities is to provide a context in which taking a decision is less hazardous as the whole information is taken into account. Most of the time, such a strategy is more efficient than making a decision on each modality and grouping the independent decision afterward. In order to do so, the classical method is to associate probability scores to each possible decision for each modality and use them as input vectors in a probabilistic inference system which fuses the pieces of knowledge under some rules expressed as conditional dependencies. Most of the time, such a framework is efficient as it corresponds to an excellent trade-off between complexity and accuracy. Nonetheless it suffers from several drawbacks. Here are few of them:

- The likelihood associated to a hypothesis is most of the time derived from a training algorithm. This guaranties a good generalization power in cases where the training data is representative.
- This likelihood is definitely derived from an objectivist point of view on probabilities, as statistical analysis of the training data are used, but probabilistic inference is deeply subjective.
- In the particular case of gesture interpretation, there is a lot of conflictive, contradictory, incomplete and uncertain knowledge, and there are other formalisms which are more adapted to this kind of situations.

Amongst all these formalisms, the one of belief function is really powerful. Moreover, it is close enough to the probabilistic formalism to keep some of its advantages and to allow an intermediate modeling where some interesting properties of both probabilities and belief functions can be used in common.

Belief functions

Originally, this theory was introduced by (Dempster 1968) throughout the study of lower and upper bound of a set of probabilities, and it was formalized by Shafer in *A Mathematical Theory of Evidence* (Shafer 1976).

In this section, we recall the main aspects of belief functions from (Shafer 1976). Let $X=\{x_1,\ldots x_M\}$ be a set of M variables and Ω_X be the set of N exhaustive and exclusive multivariate hypotheses $\{h_1, \ldots, h_N\}$ that can be associated to X. Ω_X is the **frame of discernment** (of **frame** for short) for X. Let 2^{Ω_X} be the set of all the subsets A of Ω_X, including the empty set:

$$2^{\Omega_X} = \{A / A \subseteq \Omega_X\}$$

2^{Ω_X} is called the **powerset** of Ω_X. Let m a **belief mass function** (or BF for short) over 2^{Ω_X} that represents our belief on the hypotheses of Ω_X:

$$m : \begin{cases} 2^{\Omega} \to [0,1] \\ A \mapsto m(A) \end{cases} \quad \text{with} \quad \begin{cases} \sum \left(m(A) \,|\, A \subseteq \Omega \right) = 1 \\ m(\varnothing) = 0 \end{cases}$$

$m(A)$ represents the belief that is associated exactly to A, and to nothing wider or smaller. A **focal set** is an element of 2^{Ω_X} (or a subset of Ω_X) with a non-zero belief. A **consonant BF** is a BF with nested focal sets with respect to the inclusion operator ($\subseteq$). The **cardinal** of a focal set is the number of elements of the frame it contains.

Let m be a BF over Ω_X and X and Y two sets of variables so that $X \subseteq Y$. The **vacuous extension** of m to Y, noted $m^{\uparrow Y}$ is defined so that:

$$m^{\uparrow Y}\left(A \times \Omega_{Y \setminus X}\right) = m\left(A\right) \qquad \forall A \subseteq 2^{\Omega_X}$$

Basically, it means that the vacuous extension of a BF is obtained by extending each of its focal sets by adding all the elements of Ω_Y which are not in Ω_X.

The combination of N BFs from independent sources is computed using the **Dempster's rule of combination**. It is a N-ary associative and symmetric operator, defined as follows:

$$(\cap) : \overbrace{\mathcal{B}^{\Omega_{X_1}} \times \mathcal{B}^{\Omega_{X_2}} \times \ldots \times \mathcal{B}^{\Omega_{X_N}}}^{N} \to \mathcal{B}^{\Omega_X}$$

$$m_1 \ (\cap) \ m_2 \ (\cap) \ldots (\cap) \ m_N \quad \mapsto m_{(\cap)}$$

with $\mathcal{B}^{\Omega_{Xi}}$ being the set of BFs defined on Ω_{Xi} and with Ω_X being the **cylinder product** of the Ω_{Xi}:

$$\Omega_X = \Omega_{X_1} \times \left[\Omega_{X_2} \setminus \left(\Omega_{X_1} \cap \Omega_{X_2}\right)\right] \times \ldots \times \left[\Omega_{X_{N-1}} \setminus \left(\bigcap_{i=1}^{N-1} \Omega_{X_i}\right)\right]$$

and with

$$m_{(\cap)}(A) = \frac{1}{1-\mathrm{K}} \cdot \sum \left(\prod_{n=1}^{N} m_n^{\uparrow X}(A_n) \,\middle|\, \bigcap_{n=1}^{N} A_n = A \right) \qquad \forall A \subseteq 2^{\Omega_X}$$

where the vertical bar indicating on its right the condition that A should fulfil in order to be taken account in the summation (we use this notation when the condition would be difficult to read on subscript under the summation sign). The normalizing constant

$$\mathrm{K} = \sum \left(\prod_{n=1}^{N} m_n^{\uparrow X}(A_n) \,\middle|\, \bigcap_{n=1}^{N} A_n = \varnothing \right)$$

quantifies the amount of incoherence among the BFs to fuse.

The **refinement** operation permits to express the knowledge in a more refined manner, by using a more precise frame than the one on which the original BF is defined. It is defined as follow: let two frames Ω_1 and Ω_2,

and R an application from the powerset of Ω_1 to the powerset of Ω_2, so that:

- the set $\{R(\{h\}), h \in \Omega_1\} \subseteq 2^{\Omega_2}$
- the set $\{R(\{h\}), h \in \Omega_1\}$ is a partition of Ω_2
- $\forall A_1 \subset \Omega_1, R(\{A_1\}) = \bigcup (R(\{h\}) | h \in A_1)$

BFs are also widely connected to fuzzy set theory. It appears that membership functions on Ω are included in $\mathcal{B}^{\Omega}$. Consequently, fuzzy sets are BFs and moreover, they are particularly easy to manipulate and to combine with the Dempster's rule (Dempster 1968). In that fashion, the link between the subjective part of the probabilities and the confidence measure in the fuzzy set theory is perfectly supported in the BF framework.

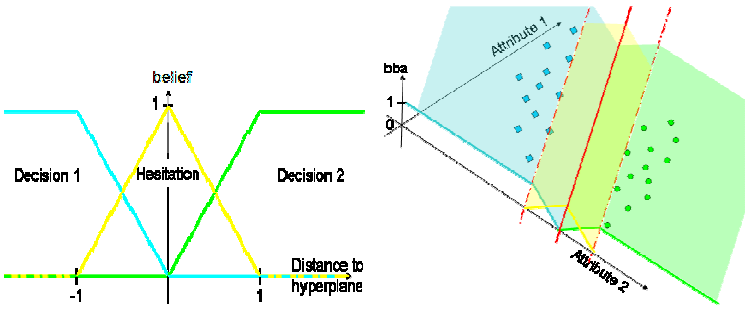


Fig. 20. Superposition of a membership function on the feature description where the SVM algorithm works.

Derivation of new belief-based tools for multimodal fusion

Evidential Combination of SVM An efficient method to solve multi-classification problems is to use a bank of binary classifiers, such as SVM) and to fuse their partial results into a single decision. We propose to do so in the BF framework. As it is proved in (Burger et al. 2006a, 2007a), the BF formalism is particularly adapted as it allows an optimal fusion of the various pieces of information from the binary classifiers. Thanks to the margin defined in SVMs, it is possible to implement the notion of hesitation easily and thus, to benefit from the richness of the BF modeling. In order to associate a BF to the SVM output, we rely on the strong connection between fuzzy sets and BFs, as explained in (a et al. 2006) and Fig. 20.

In order to make sure that the BF associated to each SVM are combinable via the Dempster's rule, it is necessary to apply a refinement from the frame of each SVM (made of two classes), to the frame of the entire set of classes, but then, it provides more accurate results than classical methods.

Evidential Combination of Heterogeneous Classifiers In case the binary classifiers involved in the process are not SVMs, then our method is not applicable anymore. As no margins are defined altogether with the separation between the classes, there is no trivial support for the hesitation distribution. An alternative is to use one of the numerous classifiers which directly provide a BF, such as CrEM (Vannoorenberghe and Smets 2005), Evidential K-NN, Expert systems, and Evidential NN (Denoeux 1995, 1997, 2000).

Another alternative is to use classical classifiers (no margins, no BF outputs), but to consider W the width of the support for the hesitation as an unknown value to determine by a learning or a cross validation.

The main interest of this evidential combination is to permit the simultaneous use of heterogeneous classifiers. As long as a classifier provides a BF, this latter can be fused with other BFs from other classifiers thanks to the conjunctive combination. This is particularly interesting when it is necessary to consider very wide sets of features which cannot be expressed in the same formalism.

Evidential Combination of Unary Classifiers It also possible to use a similar scheme (the definition of the support of the hesitation pattern via cross-validation) in order to extend the Evidential Combination of classifiers to the case of unary classifiers. In such a case, the point is to associate a generative model (without any discrimination power) to each class, to let them in competition. Each unary classifier provides a likelihood score between the generative model and the item to classify.

Then, it is possible to consider that the whole system provide an array of scores, each score being a likelihood value for each item to classify. If we assume that the highest the score, the more creditable the corresponding class (it corresponds to the first of the Cox-Jaynes axiom for the definition of subjective probabilities (Cox 1946)), then, it is possible to infer an evidential output with all the advantages it brings.

By considering the result of the algebraic comparison of the scores of each of the couple of classes, on obtains a series of values which are very similar to the precursors of the EBFs: they actually indicates the comparative membership of the item for each class of the two considered classes, in a equivalent way to a bank of SVM. The only difficulty remains to determine the values which separate the certitude of a class with respect to

another one, or on the contrary, the doubt. Here again, we propose the use of the cross-validation.

Decision Making: Partial Pignistic Transform

When a decision is to be made with respect to a peculiar problem, there are two kinds of behavior: to wait for the proofs of the trueness of one of the hypotheses, or to bet on one of them, with respect to its interest and risk. These two behaviors are considered as antagonist and it appears that no mathematical model allows making a decision which is a mix of these two stances. Consequently, we propose to generalize the Pignistic Transform, a popular method to convert BF into probabilities (Smets and Kennes 1994), in order to fill this lack (Burger and Caplier 2007).

Let γ be an **uncertainty threshold** and S^{γ} be the set of all the sets of the frame for which the cardinal is between 0 and γ (It is a truncation of the powerset to the elements of cardinal smaller than or equal to γ). We call S^{γ} the γ^{th} **frame of decision**

$$S^{\gamma} = \left\{ A \in 2^{\Omega} \;\middle|\; |A| \in \lfloor 0, \gamma \rfloor \right\}$$

where |.| is the cardinality function. The result $M_{\gamma}(.)$ of the **Partial Pignistic Transform** of order γ (noted γ^{th}-PPT) of $m(.)$ is defined on 2^{Ω} as:

$$M_{\gamma}(A) = \begin{cases} m(A) & \text{if} \quad A = \varnothing \\ m(A) + \sum \left(\dfrac{m(B) \cdot |A|}{\sum_{k=1}^{\gamma} \binom{|B|}{k} . k} \;\middle|\; \begin{matrix} B \supseteq A \\ B \notin S^{\gamma} \end{matrix} \right) & \text{if} \quad A \subseteq S^{\gamma} \\ 0 & \text{otherwise} \end{cases}$$

Then, the decision is made by simply choosing the element of the γ^{th} frame of decision which is the most believable, i.e. which gathers the highest score:

$$D^{*} = \underset{2^{\Omega}}{\operatorname{argmax}} \left(M_{\gamma} \right)$$

Application for multimodal fusion

Automatic clustering A first classical method is to use the confusion matrix of the HMM based classifier to automatically identify sign clusters. The confusion matrix is converted to a sign cluster matrix by considering the confusions for each sign. Signs that are confused form a cluster. For example, assume that sign *i* is confused with sign *j* half of the time. Then the sign cluster of class *i* is {*i*,*j*}. The sign cluster of class *j* is separately

calculated from its confusions in the estimation process. The disadvantage of this method is its sensitivity to odd mistakes which may result from the errors in the feature vector calculation as a result of bad segmentation or tracking.

We propose a more robust alternative which evaluates the decisions of the classifier and only consider the uncertainties of the classifier to form the sign clusters. For this purpose, we define a hesitation matrix. Its purpose is close to the classical confusion matrix, but it contains only the results of the uncertain decision, regardless with their correctness. Then, when a decision is certain (either true or false), it is not taken into account do define the hesitation matrix. On the contrary, when a decision is uncertain among sign *i* and sign *j*, it is counted in the hesitation matrix regardless with the ground truth of the sign being, *i*, *j* or even *k*. As a matter of fact, the confusion between a decision (partial or not) and the ground truth can be due to any other mistake (segmentation, threshold effect, etc…) whereas, on the contrary, the hesitation on the classification process only depends on the ambiguity at the level of the classification features with respect to the class borders. Then, it is more robust. In addition, it is not necessary to know the ground truth on the validation set on which the clusters are defined. This is a determining advantage in case of semi-supervised learning to adapt the system to the coder's specificity.

Partial Decision Thanks to the PPT, it is possible to make partial decisions, which is particularly adapted to classification problems where the classes are defined in a hierarchical manner (dendrogram), such as explained in (Burger and Caplier 2007), where an illustration is given on the interest of the PPT to perform automatic lip-reading on French vowels. On classical problems where such a hierarchical does not exist (such as SL recognition), it is possible to simply let it appear by defining clusters based on the hesitation matrix described above. Then, during the decision making procedure, all the pieces of information are fused together and convert into an evidential formalism via the use of the Evidential Combination. Then, the format of the result of the Evidential Combination is naturally suitable to apply the PPT.

Optional sequential decision step The only problem with such a method is that it does not guaranty that a decision is made: when the data are too uncertain, the PPT does not make any decision. Then, it can be fused with some other information, and finally, a last hesitation-free decision is taken. In (Aran et al. 2007), after a first decision step allowing some partial decisions, we propose to add some less conflictive non-manual information (that could not be taken into account earlier in the process without raising the amount of uncertainty) in order to perform a second decision step. The

originality of the method is that this second step is optional: if no hesitation occurs at the first step, the good decision is not put back into question. This is possible thanks to the use of the PPT which automatically makes the most adapted decision (certain or not). We call this original method **sequential belief-based fusion**. Its comparison with classical methods demonstrates its interest for the highly conflictive and uncertain decision required in a gesture recognition system.

Applications

Sign Language Tutoring Tool

SignTutor is an interactive platform that aims to teach the basics of sign language. The interactivity comes from the automatic evaluation of the students' signing and visual feedback and information about the goodness of the performed sign. The system works on a low-cost vision based setup, which requires a single webcam, connected to a medium-level PC or a laptop that is able to meet the 25 fps in 640x480 camera resolution requirement.

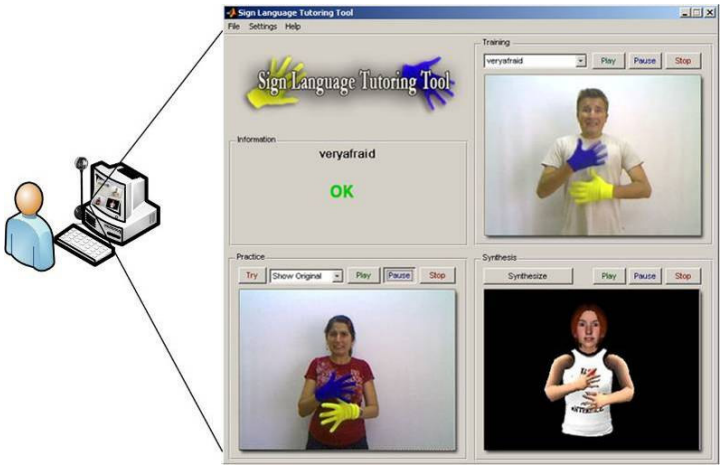


Fig. 21. SignTutor user interface

To enable the system to work in different lighting conditions and environments, the system requires the user to wear two colored gloves on each hand. With the gloves worn on the hands and no other similarly colored objects in the camera view, there are no other restrictions.

The current system consists of 19 ASL signs that include both manual and non-manual components. The graphical user interface consists of four panels: Training, Information, Practice and Synthesis (Fig. 21). The training panel involves the pre-recorded sign videos. These videos are prepared for the students' training. Once the student is ready to practice, and presses the try button, the program captures the students sign video.

The captured sign video is processed to analyze the manual and non-manual components. Here, we give a brief summary of the analysis, fusion and recognition steps. The techniques described here are also explained in the previous sections in detail so we only indicate the name of the technique and do not give the details. More details can be found in (Aran et al. 2006).

The analysis of the manual features starts with hand detection and segmentation based on the glove colors. Kalman filtering is used to smooth the hand trajectory and to estimate the velocity of each hand. The manual features consist of hand shape, position and motion features. Hand shape features are calculated from the ellipse fitted on each hand and a mask placed on the bounding box. Hand position at each frame is calculated by the distance of each hand center of mass to the face center of mass. As hand motion features, we used the continuous coordinates, and the velocity of each hand center of mass. The starting position of the hands are assumed as the (0,0) coordinate.

In this system, the head motions are analyzed as the non-manual component. The system detects rigid head motions such as head rotations and head nods with the help of retina filtering as described in the previous sections. As a result, the head analyzer provides three features per frame: the quantity of motion and the vertical, horizontal velocity.

The recognition is applied via sequential belief-based fusion of manual and non-manual signs (Aran et al 2007). The sequential fusion method is based on two different classification steps: In the first step, we perform an inter-cluster belief-based classification, using a general HMM that receives all manual and non-manual features as input in a single feature vector. A BF is derived from this bank of HMMs via the evidential combination. Then, the PPT is applied. This first step gives the final decision if there is no uncertainty at this level. Otherwise, a second optional step is applied. In this second step, we perform an intra-cluster classification and utilize the non-manual information in a dedicated model. The clusters are determined via the hesitation matrix automatically from the training set, prior to HMM training.

At the end of the sign analysis and recognition, the feedback about the students' performance is displayed in the information panel. There are three types of results: "ok" (the sign was confirmed), "false" (the sign was

wrong) and “head is ok but hands are false”. Possible errors are also shown in this field. The students can also watch a simple synthesized version of their performance on an avatar.

Cued Speech Manual Gesture Interpreter

In this chapter, we have presented several techniques in order to deal with FCS recognition:

- Hand segmentation
- Hand analysis : reduction of the variability of the shape and definition of the pointing finger
- Hand shape recognition (the shape descriptors are the FMD and the classification method is a 1vs1 Evidential Combination of SVMs followed by a PPT with an uncertainty parameter of 1 or 2)
- Face and feature detection
- Location of the pointing finger with respect to the face zones used in FCS.
- Lip segmentation
- Lip shape recognition
- Extraction of target image in case of static gestures
- Fusion of several static modalities (CS Hand shape and CS Location)

Then, the next step is to integrate all these functionalities into a global system in order to propose a French Cued Speech Translator. As the lip-reading functionality (based on the joint use of lip segmentation and lip shape recognition) as well as the fusion of manual and labial modalities (the manual gesture is static whereas the labial one is more complex (Burger 2007)) are still open issues, we propose at the moment a system which is restricted to the manual part: the CS Manual Gesture Interpreter.

This system works as follows: a CS coder (it is important to be a skilled coder, in order to produce a code in which prosody is fluent, as the dedicated retina filter is tuned for such a rhythm) wearing a thin glove of uniform but unspecified color is filmed at the frame rate of 50 images/s. The system is able to cope with unknown coder having different unknown morphology and glove. Once the video sequence is over, it is processed (this version of the interpreter works off-line), and the result is displayed. The screen is separated into two. On the left, the original video is played whereas on the right part, a virtual clone produces the gesture synchronously with the right part video (Fig. 22). Under the clone performing the recognized code, the corresponding potential phonemes are given. Note that, as no interpretation of higher level than the phonemic one is per-

formed, the system is not restricted to any dictionary, and any French message can be processed.

Fig. 22. User interface for the display of FCS manual gesture interpreter result.

Conclusion

Gestural interfaces can aid the hearing impaired to have more natural communication with either a computer or with other people. Sign language, the primary means of communication among the hearing impaired, and cued speech, which enriches lipreading with hand and facial cues, are inherently multimodal means of communication: They use gestures of the body, hands and face. Computer vision techniques to process and analyze these modalities have been presented in this chapter. These steps, as summarized below, are essential for an accurate and usable interface.

- A thorough analysis of each visual modality that is used to convey the message
- The identification of static and temporal properties of each modality and their synchronization
- Independent modeling and recognition of static/dynamic modalities
- The integration of various modalities for accurate recognition

We concentrated on sign languages and cued speech for two reasons: (1) Sign languages and cued speech are the two main media of hearing impaired communication; (2) they have different static and temporal cha-

racteristics, thus require different analysis and fusion techniques. After treating the problem in its most general form, we present two example applications: A sign language tutor that is aimed to teach signing to hearing people; and a cued speech manual gesture interpreter. The techniques discussed are general and can be used to develop other applications, either for the hearing impaired or for the general population, in a general modality replacement framework.

References

Abboud B, Davoine F, Dang M (2004) Facial expression recognition and synthesis based on appearance model. Signal Processing: Image Communication, 19(8)723-740

Adam S, Ogier JM, Cariou C, Mullot R, Gardes J, Lecourtier Y (2001) Utilisation de la transformée de Fourier-Mellin pour la reconnaissance de formes multi-orientées et multi-échelles : application à l'analyse de documents techniques". Traitement du Signal, 18(1)

Aleksic P, Williams J, Wu Z, Katsaggelos A (2002) Audio-Visual Speech Recognition using MPEG-4 Compliant Features. Eurasip Journal on Applied Signal Processing, Special Issue on Joint Audio-visual speech processing, pp.1213-1227

Aran O, Akarun L (2006) Recognizing Two Handed Gestures with Generative, Discriminative and Ensemble Methods via Fisher Kernels, Lecture Notes in Computer Science: Multimedia Content Representation, Classification and Security International Workshop, MRCS 2006, Istanbul, Turkey, pp 159-166.

Aran O, Ari I, Benoit A, Campr P, Carrillo AH, Fanard F, Akarun L, Caplier A, Rombaut M, Sankur B (2006) Sign Language Tutoring Tool. In: eNTERFACE 2006, The Summer Workshop on Multimodal Interfaces, Dubrovnik, Croatia, pp 23-33

Aran O, Burger T, Caplier A, Akarun L (2007) Sequential Belief-Based Fusion of Manual and Non-Manual Signs". Gesture Workshop, Lisbon, Portugal

Attina V (2005) La Langue française Parlée Complétée : production et perception. Thèse de Doctorat en Sciences Cognitives, Institut National Polytechnique de Grenoble, France.

Awad G, Han J, Sutherland A (2006) A Unified System for Segmentation and Tracking of Face and Hands in Sign Language Recognition. In: ICPR 06: Proceedings of the 18th International Conference on Pattern Recognition, IEEE Computer Society, Washington, DC, USA, pp 239-242

Beaumesnil B, Chaumont M, Luthon F (2006) Liptracking and MPEG4 Animation with Feedback Control. IEEE International Conference On Acoustics, Speech, and Signal Processing

Bengio Y, Frasconi P (1996) Input-output HMM's for sequence processing. IEEE Transactions on Neural Networks, 7(5):1231-1249

Benoit A, Caplier A (2005) Biological approach for head motion detection and analysis. EUSIPCO 2005, Antalya, Turkey

Benoit A, Caplier A (2005) Head Nods Analysis : Interpretation Of Non Verbal Communication Gestures. IEEE ICIP 2005, Genova

Benoit A, Caplier A (2005) Hypo-vigilance Analysis: Open or Closed Eye or Mouth? Blinking or Yawning Frequency?. IEEE AVSS 2005, Como, Italy

Black MJ, Yacoob Y (1997) Recognizing Facial Expression in Image Se-quences Using Local Parameterized Models of Image motion. International Journal of Computer Vision, 25(1):23-48

Brand M, Oliver N, Pentland A (1997) Coupled hidden Markov models for com-plex action recognition. In: IEEE Computer Society Conference on Com-puter Vision and Pattern Recognition (CVPR97), pp 994

Bullier J (2001) Integrated model of visual processing. Brain Research, 36(2-3):96-107

Burger T (2007) Reconnaissance automatique des gestes de la Langue française Parlée Complétée. Thèse de Doctorat, France

Burger T, Caplier A (2007) Partial Pignistic Transform". International Journal of Approximate Reasoning, Submitted.

Burger T, Aran O, Caplier A. (2006a) Modeling Hesitation and Conflict: A Belief-Based Approach for Multi-class Problems. In: ICMLA 06: Proceedings of the 5th International Conference on Machine Learning and Applications, IEEE Computer Society, Washington, DC, USA, pp 95-100.

Burger T, Benoit A, Caplier A (2006b) Extracting static hand gestures in dynamic context. Proceeding of ICIP'06, Atlanta, USA

Burger T, Caplier A , Perret P (2007a) Cued Speech Gesture Recognition: a First Prototype Based on Early Reduction. International Journal of Image and Vid-eo Processing, Special Issue on Image & Video Processing for Disability.

Burger T, Urankar A, Aran O, Akarun L, Caplier A. (2007b) Cued Speech Hand Shape Recognition. In:2nd International Conference on Computer Vision Theory and Applications (VISAPP07), Spain.

Caplier A, Bonnaud L, Malassiotis S, Strintzis MG (2004) Comparison of 2D and 3D analysis for automated Cued Speech gesture recognition. In:SPECOM.

Cohen I, Cozman FG, Sebe N, Cirelo MC, Huang TS (2003) Learning Bayesian network classifiers for facial expression recognition using both la-beled and unlabeled data, Proc. IEEE Computer Vision and Pattern Recognition

Cohn JF, Zlochower AJ, Lien JJ, Kanade T (1998) Feature-Point Tracking by Opt-ical Flow Discriminates Subtles Differences in Facial Expression, Proc. IEEE International Conference on Automatic Face and Gesture Recognition, April, Nara, Japan, pp. 396-401

Cootes TF, Hill A, Taylor CJ, Haslam J (1994) Use of Active Shape Models for Locating structures in Medical Images, Image and Vision Computing, 12(6):355-365

Cornett RO (1967) Cued Speech. American Annals of the Deaf 112:3-13

Cox RT (1946) Probability, Frequency, and Reasonable Expectation. American Journal hysique, 14:1-13

Dempster AP (1968) A generalization of Bayesian inference. Journal of the Royal Statistical Society, Series B, 30(2):205–247

Denoeux T (1995) A k-nearest neighbour classification rule based on Dempster-Shafer theory. IEEE Transactions on Systems, Man and Cybernetics, 25(5):804–813

Denoeux T (1997) Analysis of evidence-theoretic decision rules for pattern classification. Pattern Recognition, 30(7): 1095–1107

Denoeux T (2000) A neural network classifier based on Dempster-Shafer theory. IEEE Transactions on Systems, Man and Cybernetics A, 30(2):131–150

Essa IA, Pentland AP (1997) Coding, Analysis, Interpretation, and Recognition of Facial Expressions. IEEE Trans. on Pattern Analysis and Machine Intelligence, 19(7)757-763

Eveno N, A. Caplier A, Coulon PY (2004) Automatic and Accurate Lip Tracking. IEEE Transactions on Circuits and Systems for Video technology, 14(5)706-715

Fang G, Gao W, Zhao D (2007) Large-Vocabulary Continuous Sign Language Recognition Based on Transition-Movement Models, IEEE Transactions on Systems, Man and Cybernetics, Part A 37(1):1-9

Fasel B, Luettin J (2003) Automatic Facial Expression Analysis: A Survey. Pattern Recognition, 1(30):259-275

Feris R, Turk M, Raskar R, Tan K, Ohashi, G (2004) Exploiting Depth Discontinuities for Vision-Based Fingerspelling Recognition. In: CVPRW 04: Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW04), IEEE Computer Society, Washington, DC, USA, pp 155

Gacon P, Coulon PY, Bailly G (2005) Non-Linear Active Model for Mouth Inner and Outer Contours Detection. European Signal Processing Conference, Antalya, Turkey

Gao Y, Leung MKH, Hui SC, Tananda MW (2003) Facial Expression Recognition From LineBased Caricatures, IEEE Trans. on System Man and Cybernetics- PART A: System and Humans, 33(3)

Habili N, Lim C, Moini A (2004) Segmentation of the face and hands in sign language video sequences using color and motion cues. IEEE Trans. Circuits Syst. Video Techn. 14(8):1086-1097

Hammal Z, Couvreur L, Caplier A, Rombaut M (2007) Facial Expres-sion Classification: An Approach based on the Fusion of Facial Deformation unsing the Transferable Belief Model. Int. Jour. of Approximate Reasonning

Hérault J, Durette B (2007) Modeling Visual Perception for Image Processing. F. Sandoval et al. (Eds.): IWANN 2007, LNCS 4507, Springer-Verlag Berlin Heidelberg, pp.662–675

Hjelmäs H, Low B (2001) Face detection: a survey. Computer Vision and Image Understanding, 83:236-274

Holden E, Lee G, Owens R (2005) Australian sign language recognition, Machine Vision and Applications 16(5):312-320.

Hu MK (1962) Visual pattern recognition by moment invariants. IRE Trans. on Information Theory, 8:179-187

Kass M, Witkin A, Terzopoulos D (1988) Snakes: active contour models. Int. Journal of Computer Vision, 1(4):321-331

Keskin C, Balci K, Aran O, Sankur B, Akarun L (2007) A Multimodal 3D Healthcare Communication System. In: 3DTV Conference, Greece

Liddell SK (2003) Grammar, Gesture, and Meaning in American Sign Language, Cambridge University Press

Lien JJ, Kanade T, Cohn JF, Li C. (1998) Subtly different facial expression recognition and expression intensity estimation, Proc. IEEE Computer Vision and Pattern Recognition, Santa Barbara, CA, pp. 853-859

Luettin J, Thacker N, Beet S (1996) Statistical Lip Modeling for Visual Speech Recognition. In Proceedings of the 8th European Signal Processing Conference (Eusipco'96)

MPT:Machine Perception Toolbox, face detection algorithm: http://mplab.ucsd.edu/grants/project1/free-software/MPTWebSite/introductionframe.html

Norkin CC, Levangie PK (1992) Joint structure and function. (2nd ed.). Philadelphia: F.A. Davis.

Oliver N, Pentland A, Bérard F. (2000) LAFTER: A real-time face and tracker with facial expression recognition. Pattern Recognition, 33:1369-1382

Ong SCW, Ranganath S (2005) Automatic Sign Language Analysis: A Survey and the Future beyond Lexical Meaning., IEEE Transactions on Pat-tern Analysis and Machine Intelligence 27(6):873-891

Pantic M, Rothkrantz M. (2000) Automatic Analysis of Facial Expressions: The State of the Art. IEEE Trans. on Pattern Analysis and Machine Intelligence, 22(12)

Parton BS (2006) Sign language recognition and translation: A multidisciplined approach from the field of artificial intelligence, Journal of deaf studies and deaf education 11(1):94-101

Rabiner LR (1989) A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In:Proceedings of the IEEE, pp 257-285.

Shafer G. (1976) A Mathematical Theory of Evidence, Princeton University Press

Smets P and Kennes R (1994) The transferable belief model, Artificial Intelligence, 66(2): 91–234

Stokoe WC (1960) Sign Language Structure: An Outline of the Visual Communication Systems of the American Deaf, Studies in Linguistics: Occasional papers 8

Terzopoulos D, Waters K (1993) Analysis and Synthesis of Facial Image Sequences Using Physical and Anatomical Models. IEEE Trans. On Pattern Analysis and Machine Intelligence, 15(6):569-579

Tian Y, Kanade T, Cohn JF. (2001) Recognizing Action Units for Facial Expression Analysis, IEEE Trans. Pattern Analysis and Machine Intelligence, 23(2)97-115

Tsapatsoulis N, Karpouzis K, Stamou G, Piat F, Kollias SA (2000) A fuzzy system for emotion classification based on the MPEG-4 facial defini-tion parameter set. Proc. 10th European Signal Processing Conference, Tampere, Finland

Vannoorenberghe P and Smets P (2005) Partially Supervised Learning by a Credal EM Approach. Symbolic and Quantitative Approaches to Reasoning with Uncertainty

Viola P, Jones J (2004) Robust Real Time Face Detection. International Journal of Computer Vision, 57(2):137-154

Vogler C, Metaxas D (1999) Parallel Hidden Markov Models for American Sign Language Recognition. In: International Conference on Computer Vision, Kerkyra, Greece, pp 116-122

Wu J, Gao W (2001) The Recognition of Finger-Spelling for Chinese Sign Language. In: Gesture Workshop, pp 96-100.

Yacoob Y, Davis LS. (1996) Recognizing Human Facial Expressions from Long Image Sequences Using Optical Flow, IEEE Trans. on Pattern Analysis and Machine Intelligence, 18(6):636-642

Yang MH, Kriegman D, Ahuja N (2002) Detecting face in images: a survey. IEEE Trans on PAMI, 24(1):34-58

Zhang D, Lu G (2003) Evaluation of MPEG-7 shape descriptors against other shape descriptors. Multimedia Systems 9(1)

Zhang L (1997) Estimation of the mouth features using deformable templates. Int. Conf. on Image Processing (ICIP'97), Santa Barbara, CA, Octo-ber, pp. 328–331

Zhang Z, Lyons L, Schuster M, Akamatsu S. (1998) Comparison between Geometry-Based and Gabor Wavelets-Based Facial Expression Recognition Using Multi-Layer Perceptron. Proc. IEEE International Conference on Au-tomatic Face and Gesture Recognition, pp. 454-459