Extracting static hand gestures in dynamic context

Thomas Burger, Alexandre Benoit, Alice Caplier

EXTRACTING STATIC HAND GESTURES IN DYNAMIC CONTEXT

Thomas Burger, Alexandre Benoit** and Alice Caplier ***

*** France Telecom R&D, 28, Ch. Vieux Chêne, Meylan, France –** e-mail: ***thomas.burger@francetelecom.com***
**** LIS, 46 avenue Félix Viallet, Grenoble, France –** e-mail: ***benoit@lis.inpg.fr*** & ***caplier@lis.inpg.fr***

ABSTRACT

Cued Speech is a specific visual coding that complements oral language lip-reading, by adding static hand gestures (a static gesture can be presented on a single photograph as it contains no motion). By nature, Cued Speech is simple enough to be believed as automatically recognizable. Unfortunately, despite its static definition, fluent Cued Speech has an important dynamic dimension due to co-articulation. Hence, the reduction from a continuous Cued Speech coding stream to the corresponding discrete chain of static gestures is really an issue for automatic Cued Speech processing. We present here how the biological motion analysis method presented in [1] has been combined with a fusion strategy based on the Belief Theory in order to perform such a reduction.

Index Terms— Motion analysis, belief maintenance, visual system, video signal processing

1. INTRODUCTION

Hearing-impaired can try to guess any oral message by lip-reading. Such a task is difficult, as different phonemes correspond to identical mouth shapes. In order to improve the lip-reading efficiency, Dr. Cornett developed the Cued Speech [2]. Its purpose is to add manual gestures to lip shapes so that each phoneme has a specific visual aspect. Such a "hand & lip-reading" becomes as meaningful as the oral message.

Cued Speech is based on a syllabic coding: the message is formatted into a list of Consonant-Vowel syllable (CV). Each CV is coded with a specific gesture and combined to the corresponding lip shape, so that the whole looks unique.

A gesture contains two pieces of information: a handshape for the consonant coding and a location around the face for the vowel (fig. 1). Hand coding brings the same quantity of information than lips movement. This symmetry explains why a single gesture codes several phonemes of different lip shapes.

Each {handshape + location} is a static gesture (named a *target* gesture in the remaining of the paper): it does not contain any motion. From a strictly speaking coding point of view, a coder is supposed to perform a succession of target gestures. In real coding, the hand nevertheless moves from targets to targets (as the hand can not simply appear and disappear): transition gestures are produced but have no meaning by themselves.

As we work on automatic Cued Speech recognition for telephonic applications [3], we are interested in decoding a succession of hand gestures in real time (i.e. computationally efficiently). For such a purpose, we formulate in a first hypothesis that target gestures are sufficient to decode the hand motion and transition gestures analysis are useless to be processed (with the saving in term of computation it implies).

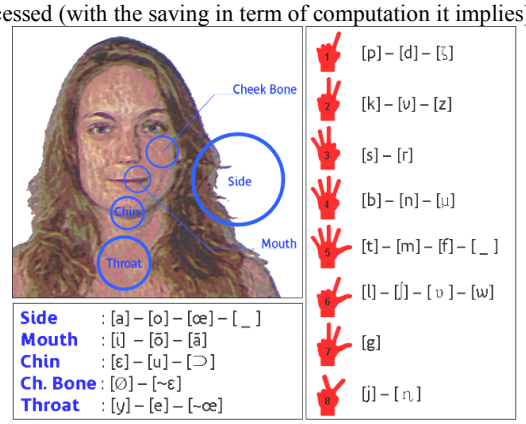

Figure 1. *French Cued Speech specifications: on the left, 5 different hand locations; on the right, 8 different handshapes.*

We also suppose in a second hypothesis, that the differentiation between transition gestures and target gestures is possible thanks to low-level kinetic information that can be extracted before the complete recognition processing of the video. We call it *early reduction* (fig. 2), in opposition to the classical late temporal integration schemes (such as HMM) which need the processing of all the images. This is motivated by the analysis of Cued Speech sequences. It shows that the hand motion is decreasing each time the hand is reaching a target. As a consequence, target gestures are related to smaller motion than transition. The main difficulty of Cued Speech hand motion analysis is that this motion is double: a global hand rigid motion associated to hand location and a local non rigid finger motion associated to hand shape formation. As a result, classical methods such as differential and block matching methods [4] or model based method [5] for hand motion analysis are not well suited.

We propose to provide the early reduction thanks to a biological approach for hand motion extraction (section 2), and a credal fusion scheme (section 3). In section 4 and 5, results are presented and discussed (as they well-ground our initial two hypotheses).

2. HAND MOTION EVALUATION

We base our analysis on [1], where algorithms simulating retina and primal visual cortex from vertebrate are described. This provides a reliable, rapid and efficient way to extract kinetic information from any video object. Our purpose is to create such a retinal processing dedicated to our specific task, i.e. hand motion evaluation.

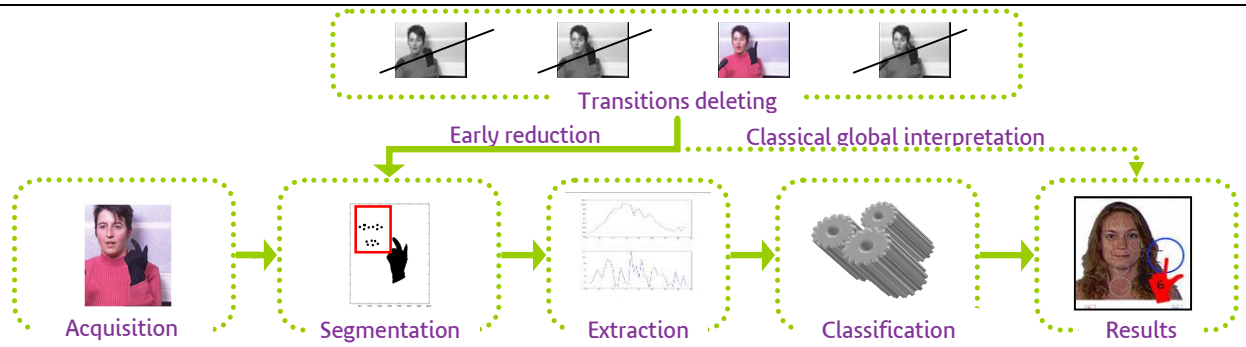

Figure 2. *Early reduction (few images processed) versus the classical late temporal integration (all the images processed)*

A *dedicated retina filter* is constituted by the following modules (fig. 3): (1) a hand detector, (2) an edges extractor and a smoothing filter, (3) an IPL filter, and (4), a sum operator.

The hand detector is a simple YCbCr auto-thresholds filter that extract the connected group of pixels which is the most likely to correspond to the coding hand (which, for segmentation and robustness reasons, wears a single color glove) [3].

The edge extractor is the filter described in [6], as the simplest robust algorithm for binary image edge detection. The smoothing filter is a 4 operations/byte approximation of a Gaussian smoother. It adapts the data to the IPL input format.

In the retina model developed in [1], the IPL filter is a biological temporal filter (its name comes from the *Inner Plexiform Layer* of the retina) which is dedicated to the analysis of moving stimuli. It can be modeled with a high pass temporal filter which enhances moving edges, particularly edges perpendicular to the motion direction. Its output can easily be interpreted in term of retinal persistence: the faster an object goes in front of the retina, the blurriest the (perpendicular to motion) edges are (fig. 3).

The sum operator integrates the output of the IPL filter in order to evaluate the "blurriness" of the edges, which can directly be interpreted as a motion energy measure. By dividing it by the edges length, we obtain a normalized measure which is homogenous to a speed measure.

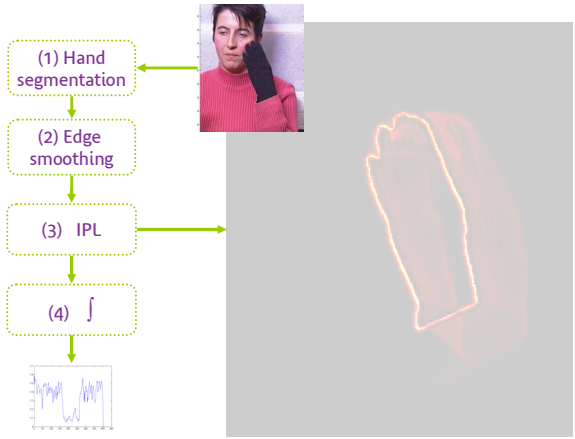

Figure 3. *A dedicated retina filter*

Cued Speech hand motion is complex, as it is a mix of a global trajectory (the movement from a location to another one) and an inner deformation (the handshape varies between gestures). Since these 2 components have different pseudo-frequencies ([7] has proved that handshape and location transitions have not the same length), we need 2 distinct filters: one for which the input is the complete image, with an IPL filter cut frequency corresponding to the slow global motion : the *Global Motion Retina Filter* (or GMRF); and another one for which the input is a vertical bounding image of the hand (to suppress global motion), with a higher cut frequency to detect high speed digit motions: the *Digit Motion Retina Filter* (or DMRF). Basically, the GMRF is accurate on hand locations targeting, whereas the DMRF is accurate on handshapes targeting. The synchronization of the outputs of the two filters and of the video stream is approximately solved by using the impulse response delays of each filter as a time offset.

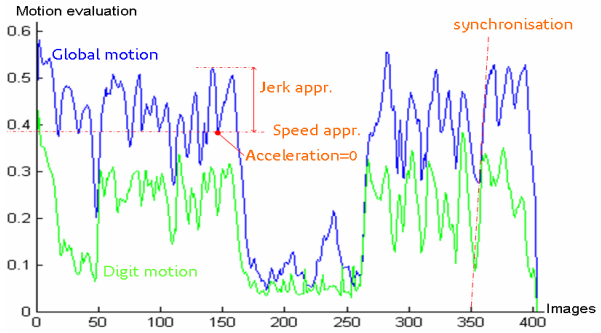

Figure 4. *GMRF and DMRF outputs*

Fig.4 presents the temporal evolution of the motion evaluation from the two dedicated retina filters during a Cued Speech sequence of 402 frames containing 19 target gestures. (Between frames 158 to 265, there is a stop in the coding process: the hand remains in the 3rd configuration without global motion). During the coding, each time a target gesture is reached, the hand speed decreases, its acceleration is null, and its jerk is negative. Moreover, the higher the absolute value of the jerk, the better realized the target: each local minimum in the motion in fig.4 is to be interpreted as a potential target. The more obvious a minimum, the better realized the corresponding target (with a strong deceleration and a long hold). Approximations of speed, acceleration and jerk are estimated for each frame and for each filter. As a result, six measures are available at each time and have to be fused in order to determine whether the current frame corresponds to a target.

3. DATA FUSION & TARGETS DETECTION

3.1. Belief Theory

Belief Theory is a convex (non-additive) generalization of probability theory. Its main advantage is to model doubt or uncertainty in a more refined manner than equi-probabilities. Originally introduced by Dempster and Shafer as a fusion method for uncertain data, it has also been compared since then to the probability theory throughout different models and interpretations [8, 9, 10].

Let **Ω** the set of *N* exhaustive and exclusive hypotheses. We call **Ω** the *frame of discernment*. Let 2^{Ω} (the *power set* of **Ω**) the set of all the subset *A* of **Ω**:

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

Let *m(.)* a bba function (a *Basic Belief Assignment* function) on 2^{Ω} that represents our belief in the propositions that correspond to the elements of 2^{Ω} :

$$m : 2^{\Omega} \to [0,1] \quad \text{with} \quad \sum_{A \subseteq \Omega} m(A) = 1$$
$$A \ \alpha \ m(A)$$

Note that:

- Contrarily to probabilistic models, the belief can be assigned to non-singleton propositions, which allows to model doubts or indecisions between elements.
- Ø belongs to 2^{Ω}. Hence, it is possible to assume an undefined hypothesis of the frame of discernment.

In our case, **Ω** = {*Target*, *Transition*} and 2^{Ω} = {Ø (i.e. *something else*), *Target*, *Transition*, {*Target* ∪ *Transition*} (i.e. *Doubt*)}. Moreover we have 6 measures coming from GMRF and DMRF filters (speed, acceleration and jerk measures) on which a bba on 2^{Ω} will be assigned.

3.2. Fuzzy sets

Through each measure we have a partial belief on the current state on which we want to make an hypothesis belonging to 2^{Ω}. Hence, we have to model the knowledge we have on each measure to associate each value it can take to a bba.

We define the bba on each of the 6 measures by modelling them through fuzzy sets: they allow modeling the balance between the belief we have on a hypothesis, and the belief we have on a wider hypothesis. Thus, it authorizes imprecise knowledge to be taken into account, while avoiding a statistical inference on the data (which prime goal is to refine the doubt or the lack of knowledge we have on a stochastic phenomenon).

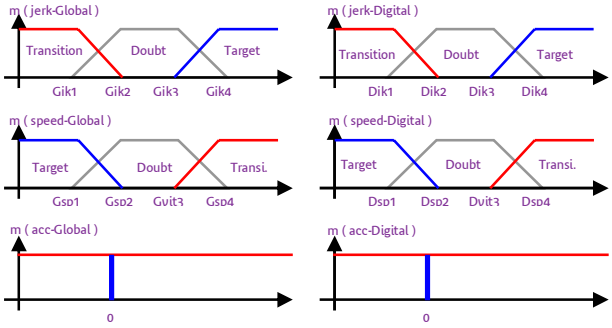

Figure 5. *bba definition for the 6 measures*

As our modelling is quite simple, we assume that all the possible hypotheses are in the power set. Hence, a null bba will be assigned to Ø. Moreover, a binary assumption is made on the information about the accelerations: if they are not null, the corresponding gesture is a transition gesture. For the remaining measures, fuzzy sets are defined as trapezoids (fig. 5). One nevertheless needs to define 16 values for the smooth thresholds. The interest of the method is the high robustness to these value variations, as they are by nature imprecise. Hence, they are semi-automatically defined.

The assumption on the acceleration and the original knowledge on which thresholds are semi-automatically defined, are both based on Cued Speech video expertise.

3.3. Fusion process

The purpose is to combine (under associativity and symmetry assumptions) some bbas into a new bba, which can be done thanks to the *conjunctive combination* (which correspond to the Belief Theory extension of the AND operator). For *N* bbas $m_1, \ldots, m_N$ from N independent sources,

$$m_{\cap} = m_1 \cap m_2 \cap \ldots \cap m_N$$

$m_{\cap}$ is defined as,

$$m_{\cap} : (2^{\Omega})^N \to [0,1]$$
$$m_{\cap}(A) = \sum_{A = A_1 \cap \ldots \cap A_N} \left(\prod_{n=1}^{N} m_n(A_n) \right) \qquad \forall A \subseteq 2^{\Omega}$$

In our application, this complex computation is obvious, as there are only 3 elements in 2^{Ω}. In the case of 2 bbas m_1 and m_2, it can be summarized in a double entry table. Each cell of the table contains the product of the 2 bbas corresponding to the entries (tab. 1). The summation of all the cells with respect to a particular pattern gives the combined bba on each element of 2^{Ω}: target (pale taint), doubt (medium taint), transition (dark taint) and Ø (black).

$m_1 \setminus m_2$	**Target**	**Doubt**	**Transition**
Target	m1(ta). m2(ta)	m1(ta). m2(d)	m1(ta). m2(tr)
Doubt	m1(d). m2(ta)	m1(d). m2(d)	m1(d). m2(tr)
Transit.	m1(tr). m2(ta)	m1(tr). m2(d)	m1(tr). m2(tr)

Table 1. *The conjunctive combination of 2 sources*

One can also define one's own combination rule by defining the way the cells are summed in such a table (as long as it remains associative and symmetric). For example, the tab. 2 corresponds to another fusion scheme, the *disjunctive combination* (corresponding to the extension of the OR operator and for which the ∩ operator is replaced by a ∪ operator).

$m_1 \setminus m_2$	**Target**	**Doubt**	**Transition**
Target	m1(ta).m2(ta)	m1(ta). m2(d)	m1(ta). m2(tr)
Doubt	m1(d). m2(ta)	m1(d). m2(d)	m1(d). m2(tr)
Transit.	m1(tr). m2(ta)	m1(tr). m2(d)	m1(tr). m2(tr)

Table 2. *The disjunctive combination of 2 sources*

3.4. Implementation

We decide to split the fusion into a conditional tree (fig. 6): the result of the conjunctive fusions on the GMRF data, and on the DMRF data are expressed with 2 bbas that are in turn fused (with a disjunctive combination) in a final stage.

This arborescence has two advantages: (1) it corresponds to the conditional way humans perform the integration of too big a set of data (by using intermediate fusions), and then can better

fit an expert system. (2) Each fusion node can be modified to fit minor input variations without disturbing the whole process. Hence, with another hand motion (different than Cue Speech) one might tune the final fusion process by choosing between conjunctive combination, disjunctive combination or another fusion table, depending on the desired flexibility.

Finally, the fusion output is modeled by a trivial determinist automat (under the Markov property) in order to avoid consecutive redundancy.

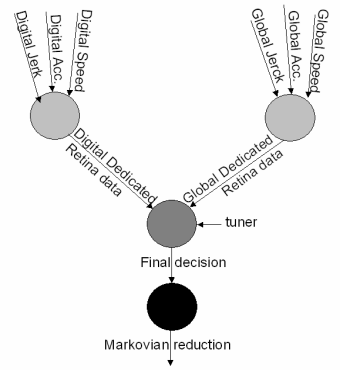

Figure 6. *The data fusion pattern*

4. RESULTS

Fig. 7 displays the values m(*target*), (which is the bba corresponding to the belief that the current image is a target), at each level of the fusion process for a sequence. On the final stage, only the images on which the bba values 1 are classified as "targets", and are transmitted to the recognition module. The remaining images are discarded. This corresponds to an approximate 90-95% (depending on the Cued Speech video) reduction of the amount of images to process. As 98% of the video corresponds to transition gestures, our reduction is not totally optimal: some transition gestures are classified as target gestures. However the reverse error (target taken for transition, which is impossible to compensate), occurred in less than 5% of case: the study of such errors have shown to be mainly originated from a bad coding to the extend that the co-articulation was too strong: human expertise has extracted such targets thanks to high-level and contextual analysis, but we do not expect our system to deal with them.

The last advantage is that our system considerably enforces the recognition stage, as handshapes and locations are easier to classify on fully realized gestures than during transitions.

This was implemented in MatLab/C on a Windows workstation. It takes 1.2s/image for targets complete recognition process [3] versus 0.3s/image for transitions discarding process. Hence, the early reduction speeds up a global sequence processing 3.6 times (4 times on 90% of the images).

5. CONCLUSION

We provide a reliable tool for reducing the number of images to process in a Cued Speech video (for recognition tasks). This tool is based on dedicated retinal algorithms and a data fusion on a belief network. Our next works will support (1) retinal outputs and synchronization refinement, and (2) generalization to various motion processing.

REFERENCES

[1] A. Benoit, A. Caplier, “Motion Estimator Inspired from Biological Model for Head Motion Interpretation”, *WIAMIS05*, Montreux, Suisse, April 2005.

[2] R. O. Cornett, "Cued Speech", *Am. Ann. Deaf*, 1967.

[3] T. Burger and al. "Cued Speech Hand Gesture Recognition Tool", *EUSIPCO'05*, Antalya, Turkey – 4-8 sept. 2005.

[4] L. Barron and al. “Performance of optical flow techniques”. *International Journal of Computer Vision*, 12(1):43-77, 1994.

[5] M. Irani and al. “Computing Occluding and Transparent Motions” *Inter. Journal of Computer Vision*, 12:1, 5-16, 1994.

[6] S. Wang and al. "Simplest operator based edge detection of binary image" *Inter. Comp. Cong.*, Logistical Engineering University, P.R.China, 28 - 30 May 2004.

[7] V. Attina, and al. "A pilot study of temporal organization in Cued Speech production of French syllables: Rules for a Cued Speech synthesizer." *Speech Communication*, vol. 44, pp. 197-214, 2004.

[8] J.M. Nigro and M. Rombaut. “Idres: a rule-based system for driving situation recognition with uncertainty management”, *Information Fusion*, dec. 2003. Vol. 4.

[9] P. Smets and R. Kennes. "The transferable belief model". *Artificial Intelligence*, 66(2): 191–234, 1994.

[10] R. Shafer and P. P. Shenoy, “Probability propagation,” *Ann. Math. Art. Intel.*, vol. 2, pp. 327–352, 1990.

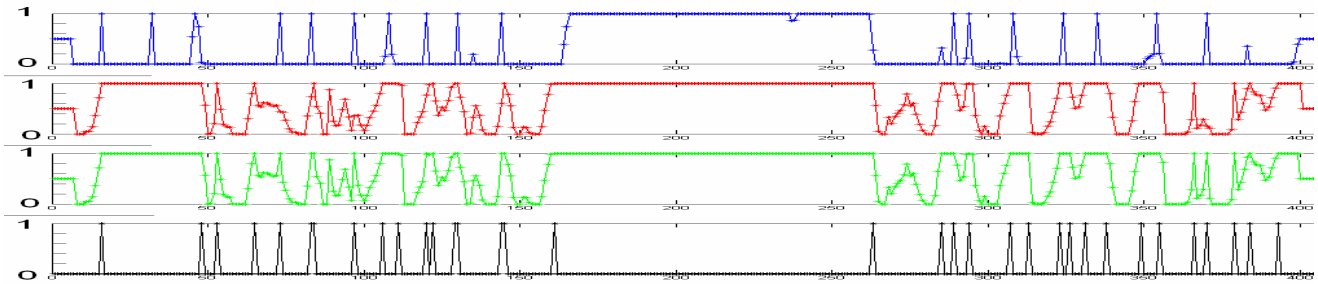

Figure 7. *Temporal evolution of m(target) for a video sequence of 402 images after the fusions of* ***(1)*** *GMRF and,* ***(2)*** *DMRF measures,* ***(3)*** *final fusion,* ***(4)*** *Markovian reduction. At the end of the process, only 34 images remain believed as targets.*