Distributed data fusion applied to human gesture measurement

Eric Benoit, Didier Coquin, Hideyuki Sawada

Distributed data fusion applied to human gesture measurement.

E. Benoit (1), D. Coquin (1), H. Sawada (2)

(1) LISTIC-ESIA, Université de Savoie,
B.P. 806, F-74016 Annecy cedex, France
Tel: +33 450 096 540, Fax: +33 450 096 559 E-mail: didier.coquin, eric.benoit @univ-savoie.fr
(2) Kagawa University
2217-20 Hayashi-cho, Takamatsu, Kagawa, 761-0396 JAPAN
E-mail: sawada@eng.kagawa-u.ac.jp

Abstract - In order to improve the link between an operator and its machine, some human oriented communication systems are now using natural languages like speech or gesture. In this paper, a gesture recognition system based on the fusion of measurements issued from different kind of sources is presented. Sources are a dataglove measuring the hand gestures and a video camera measuring the general arm gestures. The measurements used are partially complementary and partially redundant. The application is distributed on intelligent cooperating sensors. The paper presents the measurement of the hand and the arm gestures, the fusion processes, and the implementation solution.

I. INTRODUCTION

Gesture measurement is a complex process that needs to be performed with several sensor technologies in order to improve the final result. In many approaches, a data glove is used to give data about the hand gesture but it cannot access to the hand position or the hand orientation. It is then necessary to add another sensors that are able to capture at least the position and the orientation of the hand such as 6Dof (6 degree of freedom) sensors or video cameras. Furthermore, the video camera is able to give some pieces of information about the arm motion. This additional data is useful to recognize a gesture.

Increasing the number of sources can be an improvement only if the measurement results can be combined. In order to simplify the fusion process, all measurements are expressed as fuzzy descriptions or fuzzy distributions of possibility. this common representation allows to combine etherogenous sources as shown in [Benoit et al., 2003] and [Sawada et al., 2004].

The fusion of the redundant information is performed by a possibilistic aggregation of measurements. This approach allows one to have sources with different confidences. The fusion of complementary information is performed by an aggregation based on fuzzy rules. This general approach based on a fuzzy representation of measurements permits the distribution of the fusion process on a distributed network of sensors.

This paper first presents the definition of the fuzzy description of a measurement and its application to the fuzzy representation of hand and arm gesture. Then the fusion process based on complementary information and redundant information is presented. Finally, an application to the driving of a mobile robot is shown.

II. MEASUREMENT

A. Fuzzy description of measurement

The link between a physical state and its linguistic representation is characterized by a symbolism defined by the triplet $<X,L,R>$ where X is the set of physical states, L is the lexical set used to represent measurement results and R is a relation on XxL. Two mappings can be extracted from this relation: The *description mapping* denoted d associates a subset of L to any item of X, and the *meaning mapping* denoted m associates a subset of X to any item of L. This two mapping are linked with the following equation.

$$\forall x \in X, \forall a \in L, x \in m(a) \Leftrightarrow a \in d(x) \quad (1)$$

The R relation can be a fuzzy relation. Then, the translation of a physical state into its linguistic representation is called a *fuzzy linguistic description mapping* or simply a *fuzzy description mapping*. It transforms an object x of the set of physical states X into a fuzzy subset of linguistic terms called the *fuzzy description* of x. The dual mapping, called the *fuzzy meaning* mapping, associates a fuzzy subset of X to each term l of the lexical Set L. This fuzzy subset is the *fuzzy meaning* of l. This two mapping are linked with the following equation:

$$\forall x \in X, \forall a \in L, (\mu_{m(a)}(x) = \mu_{d(x)}(a)) \quad (2)$$

B. Hand gestures

A fuzzy description of hand postures add been proposed in [Allevard et al. 2003]. Like in this paper, fingers configurations are given numerically by the Cyberglove©'s bending sensors.

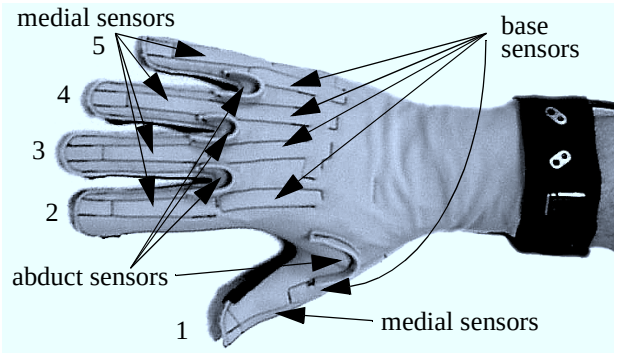

Fig. 1 Picture of the Cyberglove, name of the sensors.

The numerical to lexical conversion is made using a fuzzy partition of the dataglove measurement spaces. The fuzzy glove then provides a fuzzy description of each finger configuration and their relative position. The robustness and efficiency of the recognition system depends on the definition of the partitions. Different methods can be used for its construction (empirically, by interpolation, by clustering methods, etc.). The calibration of the fuzzy glove can be made by modifying this partition.

A preliminary study has shown that each finger can take five different configurations shown in Fig. 2 Depending on the complexity of the sign language to be recognized, either those five linguistic values are used or only the two simple values *bent* and *straight*. The relative finger positions can be either *separated* or *together*.

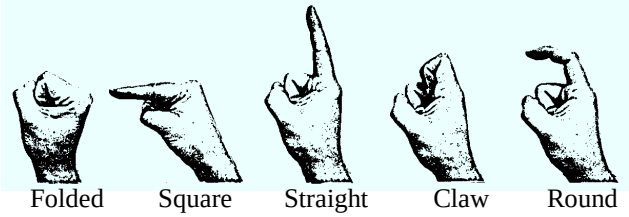

Fig. 2 Picture of the five key configurations of a finger.

To qualify linguistically the configuration of a finger, two sensors are used: the finger base sensor and finger medial sensor. Two monodimensional partitions have then to be defined on the numerical spaces of those two sensors. The corresponding lexical sets are respectively: $L_{BaseIndex}$ = {*Folded*, *Straight*} and $L_{MedialIndex}$ = {*Folded*, *Half*, *Straight*}. The final lexical set used to describe the whole index finger is: L_{Index} = {*Folded*, *Straight*, *Square*, *Round*, *Claw*}. The set of rules is given in the table of Fig. 3

Medial Sensor of index	Base sensor of index: *Straight*	Base sensor of index: *Folded*
Folded	*Claw*	*Folded*
Half	*Round*	*Round*
Straight	*Straight*	*Square*

Fig. 3 Set of rules for Index

The recognition of the hand postures is performed by using fuzzy rules which express human knowledge about how a sign is performed. For example, the sign *pointing* will be described by: *index straight and other finger bent; all fingers together*. Such a recognition system is readable and understandable. A new gesture can be added to the vocabulary very simply by giving its linguistic description or by asking the system to learn this description from a small set of examples.

The final lexical set can be small. for example we had use:

$$L=\{\textbf{\textit{flat}}, \textbf{\textit{ok}}, \textbf{\textit{thumbUp}}, \textbf{\textit{folded}}, \textbf{\textit{pointing}}, \textbf{\textit{cut}}\}. \quad (3)$$

With this lexical set, a representation of a hand posture can be the lexical fuzzy subset:

$$d(x) = \{0/\textbf{\textit{flat}}, 0/\textbf{\textit{ok}}, 0.9/\textbf{\textit{thumbUp}}, 0.1/\textbf{\textit{folded}}, 0/\textbf{\textit{pointing}}, 0/\textbf{\textit{cut}}\}. \quad (4)$$

C. Arm gestures

The recognition of arm gesture (Fig. 4) is performed with a digital video camera. The general process of the method is composed of four main parts: a preprocessing step, a feature extraction step, a training step and a classification step.

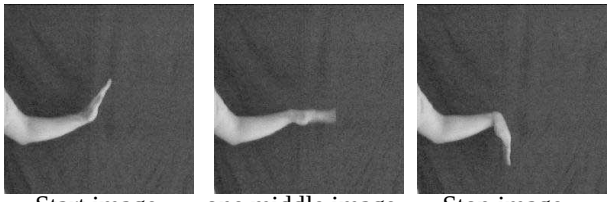

Fig. 4 Samples of an arm gesture sequence that means ***stop***.

The preprocessing step focuses on the gesture. The preprocessing chain consists in the following operations: binary image computation, binary image enhancement, and hand region extraction.

The feature extraction step uses two techniques. The first one, dedicated to static recognition, uses histogram orientation in the gesture image (by computing the local gradient on the image). It is based on an algorithm proposed in [Freeman and Roth, 1995]. The orientation histogram will act as feature set of the gesture. The second one, used for dynamic gesture recognition, extracts *the dynamic signature* and is defined using a new original method which gives a compact and efficient representation of the gesture. The dynamic signature is a binary image which represents the superposition of all the skeletons of the hand region for all images (hand postures) within the gesture image sequence (see Fig. 5). It reflects the motion information of the gesture and also the hand spatial configuration during the gesture motion. Using the proposed dynamic signature, the gesture can be segmented

into its constitutive steps such as the preparation and the stroke [Quek, 1994]. Each gesture will be represented by a particular dynamic signature.

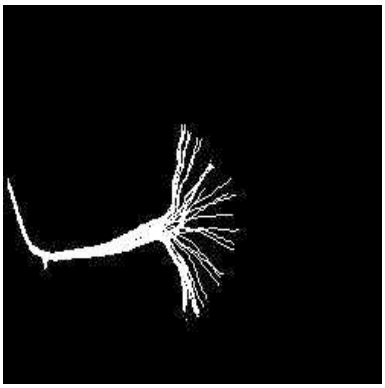

Fig. 5 Dynamic signature

In order to perform the gesture classification, the gesture alphabet has to be known. A training stage is required. Using a set of samples for each gesture, a training set of data is generated for both the static and the dynamic recognition. Using different representations in space of the same gesture gives invariance to small translations and rotations. the training set consists of *m* training sequences (dynamic signatures) for each hand gesture within the gesture alphabet. Using *n* gestures in the gesture alphabet and having *m* sets of example signatures for each one, there are *n.m* sample signatures for the training sequence.

The unknown gesture features are be compared with the features stored in the training data. The best match leads to the recognition of the unknown gesture. To compare two dynamic signatures, the Baddeley's distance is computed within the smallest frame containing the skeletons [Coquin and Bolon, 2001].

Baddeley distance [Baddeley, 1992] is defined between two binary images *A* and *B* having the same support *S*, by:

$$D(A,B) = \left[\frac{1}{card(S)}\sum_{s\in S}\left|d_A(s)-d_B(s)\right|^E\right]^{\frac{1}{E}} \quad (5)$$

with $s = (x,y)$ a pixel, exponent $E = 2$, and $d_A(s)$ is the shortest distance between s and the skeleton. $Card(S)$ is the number of element of support *S*.

The fuzzy description is obtained by the translation of the distance betweem the unknown gesture features and the features stored in the training data.

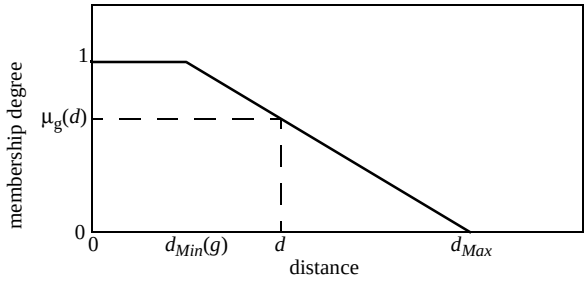

Fig. 6 Fuzzification of the distance between the dynamic signatures of the measured gesture and a training gesture *g*.

A membership function is defined for each gesture. The $d_{Min}(g)$ parameter is computed as the max distance between the items of the set of samples signatures of the *g* gesture. The d_{Max} parameter is defined as the max distance between the items of the set of samples signatures.

The lexical set used to describe arm gesture include both static and dynamic gestures:

L = {***goesBack***, ***goesAhead***, ***goesUp***, ***goesDown***, ***shakesHand***, ***staysUp***}. (6)

III. FUSION PROCESS

A. Fusion of complementary information

Sources are considered as complementary when the pieces of information concern different physical phenomenons. In this case, measurements from each sources are described separately with a symbolism $<X_i,L_i,R_i>$ for each source *i*. Each lexical sets is specific to each sensor. Then fuzzy descriptions are fused using a fuzzy aggregation based on a set of fuzzy rules.

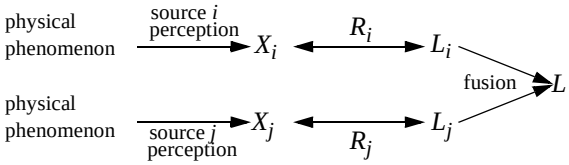

Fig. 7 Relations between sets with complementary sources.

The final lexical set is defined to represent a full gesture. A set of term used for mobile robot control are chosen.

L={***stop***, ***ahead***, ***right***, ***left***, ***behind, unknown***} (7)

The ***unknown*** term is added for all combination of hand posture and arm gesture that can not be used. For example, when the hand is ***folded*** and the arm is ***goesUp***, the resulting gesture is not use to control the robot. An other approach will be to find a term for all gestures. The term that describe last gesture will be ***rebelSign***.

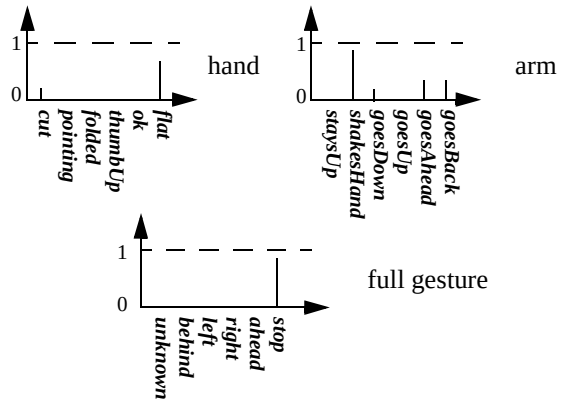

Fig. 8 Example of fusion with complementary sources

B. Fusion of redundant information

Sources are considered as redundant when the pieces of information concern the same physical phenomenons. Each measurement process is based on a symbolism $<X_i,L_i,R_i>$. When sources are partially redundant, it is supposed that the X_i sets intersect enough to be described into a unique lexical set L. In this case, the descriptions issued from the different sources are represented by fuzzy subsets of the same lexical set. In our application, the chosen lexical set is the one that can represent the full gesture (see Eq. (7) and Fig. 10). As the lexical set is not specific to a source, 2 different physical states can be described by a source with the same term. For example, in the ***stop*** the ***right*** and the ***left*** gesture, the hand is kept flat. Then hand posture is the same and can not be distinguished with the dataglove source.

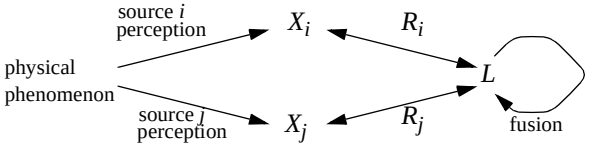

Fig. 9 Relations between sets when sources are redundant.

Considering the semantic of the membership degree $\mu_{d(x)}(a)$ of a term a to the fuzzy description $d(x)$, it can be considered as the possibility degree that a represents the state x. The representation processes do not warranty the fuzzy descriptions to be possibility distributions. Generally the fuzzy descriptions are not normalized to 1. A solution is to normalize each description. An other one is to accept non normalized distributions to represent unreliable descriptions.

Considering two symbolisms $<X_1,L,R_1>$ and $<X_2,L,R_2>$, and the associated lexical descriptions d_1 and d_2 of the same physical entity. The aggregation of the fuzzy descriptions computed with these two symbolisms can be performed with a triangular norm i.e. a Tnorm T:

$$\mu_d(a) = \mu_{d1}(a) \, T \, \mu_{d2}(a) \qquad (8)$$

This simple approach gives good results when both sources are reliable, because in this case both sources are agree to describe a gesture at least one common term.

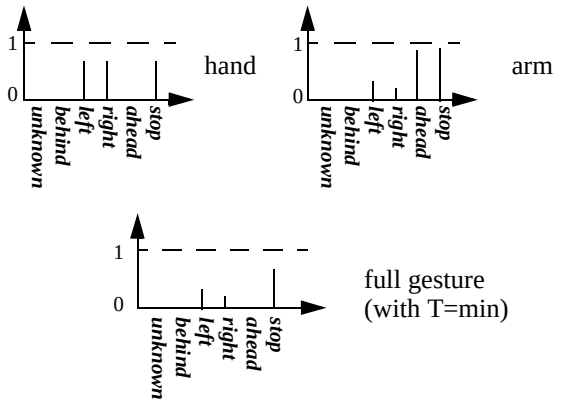

Fig. 10 Example of fusion with redundant sources

In a most general case the reliability of sources are not equivalent. The possibility theory proposes a solution to take this difference into account as for example in [Desachy et al., 1996]. To know how two sources must be fuzzed, it's only needed to know wich is more reliable. Consider d_1 as more reliable than d_2. First a consistency degree between sources is computed:

$$h(d_1, d_2) = sup_{a \in L}(min(\mu_{d_1}(a), \mu_{d_2}(a))) \qquad (9)$$

then the description of the gesture is defined by:

$$\mu_d(a) = min(\mu_{d_1}(a), max(\mu_{d_2}(a), 1\text{-}h(d_1, d_2))) \qquad (10)$$

When both sources are reliable $h(d_1, d_2) = 1$ and Eq. 10 is equivalent to Eq. 8 with T = min. Depending on the fusion to perform, other fusion processes are proposed in [Dubois and Prade, 1992] and [Oussalah, 2000].

IV. IMPLEMENTATION

This approach is applied to the control of a mobile robot with gestures. The sensors are connected to different computing systems and these systems are linked to an ethernet network and communicate with the UDP/IP protocol. The data glove is connected to a Jstik© board that is able to run java J2ME-Cldc programs. The robot is connected to another Jstik© board. The video camera is connected to a computer. A separate board runs the fusion program.

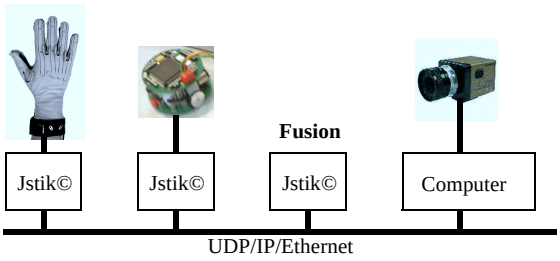

Fig. 11 Architecture of the application.

Each instrument runs as an UDP server and sends its data with a broadcast. Each frame includes the data and source identification, the indentification of the lexical set, and the data values i.e. a vector of membership degrees. The fusion board catches the frames emitted by the glove and the camera and send the result on the network. The robot catches the frames emitted by the fusion board.

The fusion board is able to select its mode between the complementary one and the redundant one. If the same lexical set is used by all sources, it switches into the redundant mode. If lexical sets are different and if it includes a set of rules using these lexical sets, it switches into the complementary mode. In the complementary mode, adding a new source need to create a new set rules to take this one into account. In the redundant mode, adding a new source is taken into account by an extension of Eq. 10 to *n* sources. In this last case the knowledge is distributed over the set of sensors. Indeed, each sensor gives a fuzzy subset that depends on the final application.

V. CONCLUSION

In the domain of gesture recognition with multiple sources of measurements, the nature of sources can be very different and may be difficult to combine. The fuzzy subset theory and the possibility theory give an framework for the aggregation of etherogeneous data. Futermore it allows to take into account the difference of reliability between sources. This paper shows how pieces of information from various sources can be fuzed considering their complementarity or their redundance.

REFERENCES

[Allevard et al., 2003] T. Allevard, E. Benoit, Foulloy L., "Fuzzy Glove for Gesture Recognition", 17th IMEKO World Congress, Dubrovnik, Croatia, pp. 2026-2031, June 2003.

[Baddeley, 1992] A.J. Baddeley, "An error metric for binary images", Robust Computer Vision, Wichmann, Karlsruhe, pp. 59-78, 1992.

[Benoit et al., 2003] E. Benoit, T. Allevard, T. Ukegawa and H. Sawada, "Fuzzy Sensor for Gesture Recognition Based on Shape Recognition of Hand", Int. Symp. on Virtual Environements, Human-Computer Interfaces, and Measurement Systems (VECIMS'03), Lugnano, Switzerland, pp. 63-67, July 2003.

[Coquin and Bolon, 2001] D. Coquin, P. Bolon, "Applications of Baddeley's distance to dissimilarity measurement between gray scale images", Pattern Recognition Letters, Vol. 22, pp. 1483-1502, 2001.

[Desachy et al., 1996], J. Desachy, L. Roux and El-H Zahzah, "Numeric and symbolic data fusion: A soft computing approach to remote sensing images analysis", Pattern Recognition Letters, Volume 17, Issue 13, pp 1361-1378, November 1996.

[Dubois and Prade, 1992] D. Dubois and H. Prade, "Combination of Fuzzy Information in the Framework of Possibility Theory", M.A. Abid, R.C. Gonzalez (Eds.), Data Fusion in Robotics and Machine Intelligence, Academic Press, New York, 1992.

[Freeman and Roth, 1995] W. T. Freeman and M. Roth, "Orientation Histograms for Hand Gesture Recognition", Mitsubishi Electric Research Laboratories, Cambridge Research centre, TR-94-03a, December 1995.

[Oussalah, 2000] M. Oussalah, "Study of some algebraical properties of adaptive combination rules", Fuzzy Sets and Systems, Volume 114, Issue 3, pp 391-409, September 2000.

[Quek, 1994] F. Quek, "Toward a vision-based hand gesture interface", Proceeding of Virtual Reality Software and Technology Conference, pp. 17-29, Singapore, August 1994.

[Sawada et al., 2004] H.Sawada, T. Ukegawa, E. Benoit, "Robust gesture recognition by possibilistic approach based on data resampling", Fuzzy Systems & Innovational Computing (FIC2004), Kitakyushu, Japan, pp. 168-173, June 2004.