Fusion method choice driven by user constraints in a human gesture recognition context

Stéphane Perrin, Eric Benoit, Didier Coquin

Fusion method choice driven by user constraints in a human gesture recognition context

Stéphane Perrin, Eric Benoit, Didier Coquin
Lab. LISTIC
Université Savoie Mont Blanc
Annecy, France
{stephane.perrin, eric.benoit, didier.coquin}@univ-smb.fr

***Abstract*—Recognition of human gesture leads to a dynamic field that produces many various methods. The main way to improve the recognition process is to perform data fusion based on a qualification of each recognition method. The advance in data fusion gives also several solutions and the choice of a fusion method is a crucial point. The goal of this paper is to present an approach were the choice of the fusion method is driven by the constraints of the final user of data. Such constraints are expressed as bounds on indicators like confidence or accuracy. A practical experimentation illustrating this approach is presented in this paper. From a depth camera sensor, human gesture is interpreted and, fusion method is selected in accordance with the constraints of the final information consumer. This approach is illustrated in a robotic control by human gesture: outputs perform the action to be done by a NAO robot.**

Keywords—gesture recognition, fusion, fuzzy logic, user constraint

I. INTRODUCTION

Human recognition of human gesture leads to a dynamic field that produces many various methods [1]. These methods commonly perform a measurement process in order to convert the physical state of an entity into an information entity. If numerical values are commonly used to represent measurement results, it is now admitted that some applications manipulate symbolic values or linguistic terms better.

In this paper, we concentrate first on recognizing human gesture from upper limb postures detection. The two parts used of upper limb, i.e. arm and forearm, is expressed using linguistic terms. The corresponding symbolic values result from numeric angles measurements of human body. These angles measurements provided by a depth camera sensor, in our case an "xtion pro live" from Asus. We use yet linguistic terms to express a posture from symbolic values of upper limb parts.

Then we focus on decision process performed from symbolic values. Symbolic values are commonly expressed in terms of mass. Fusion methods are particularly effective and used to manipulate mass [2]. Commonly used in decision system, fusion methods are able to formalized decision criteria from fuzzy information, especially when evidence theory is used. Finally we propose to express the needs of the information consumer which is view as a constraint on the decision process. This allows to choose and to compute the decision, i.e. action to do, in accordance with the consumer's needs and usage.

Gesture recognition done by a human is complex and the decision process can take several criteria in consideration: hesitation between several known gestures, believe in a recognition gesture, etc. Fuzzy logic facilitates the manipulation of these concepts. In the case of automatic and artificial process for the recognition of human gestures, we show that results of fusion methods associated to the evidence theory are a good way to express the consumer criteria. These criteria are finally used to select method fusion.

This approach, illustrated into section 5, is used to give orders to a humanoid NAO robot developed by the Aldebaran Company, from human gesture recognition. The system learns a set of gestures. Each known gesture corresponds to an order to be transmitted to the robot. The decision system chooses the order to be transmitted to the robot. For example, the user wants the robot to execute an order if and only if the recognition of the corresponding human gesture is sure, i.e. with low uncertainty or not ambiguous with other possible gestures. Another example: a gesture corresponds to an emergency action. In this case, if this gesture is recognized with enough certainty the corresponding order is chosen even if another gesture is recognized at the same time but with a greater certainty.

In the next section, we detail gesture modeling using fuzzy representation. The gesture recognition process is presented in section 3. In section 4 we introduce a fusion method in order to express the decision criteria using evidence theory. This theory allows defining and formalizing user constraints: recognition quality, threshold of gesture detection. Facilitating threshold determination based on mass transfers inside fuzzy representation of human gesture is discussed in section 5.

User constraints are thrown back to the decision system which is now able to choice robot order to be applied, in accordance to the user expectation.

II. FUZZY LOGIC FOR HUMAN GESTURE MODELING

A. *Fuzzy representation : fuzzy nominal scale*

This section reminds the concepts of fuzzy nominal scale and metrical scale presented in [3] and that are used to represent gestures. For the sake of simplicity, the notation of

any fuzzy subset u will also denote the associated membership function. Then the membership degree of a value x to a fuzzy subset u is denoted $u(x)$. The notation of the membership of a value to a function $M(s)$ is then denoted $M(s)(x)$.

Let X be a set of object states. In order to characterize linguistically any measurement over X, let L be a set of linguistic terms, representative of the physical phenomenon. Let $F(L)$ be the set of fuzzy subsets of L.

The fuzzy scales used in our study are defined by an id-symbolism $<X, L, R>$ where:

X is the set of quantity manifestations.

L refers to a set of terms dedicated to the expression of measurement values, it is often known as the universe of discourse.

R, called the representation, is a fuzzy mapping from X to L.

The conversion of a physical state into its linguistic representation is called a fuzzy linguistic description mapping [4][5] or simply a fuzzy description mapping. It transforms an object x of X into a fuzzy subset of linguistic terms $D(x) \in F(L)$ called the *fuzzy description* of x. The fuzzy subsets of linguistic terms are called *lexical fuzzy subsets* (LFS). A 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 $M(l)$ is the *fuzzy meaning* of l (see fig. 1). With fuzzy nominal scales, the fuzzy meanings are chosen to respect some characteristics of fuzzy sensors and they form a strict fuzzy partition of X.

Considering the subset $LF(L)$ of $F(L)$ defined by

$$A \in LF(L) \Leftrightarrow \sum_{l \in L} A(l) = 1 \qquad (1)$$

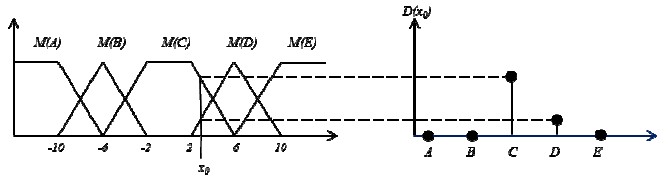

Fig. 1. An example of fuzzy nominal scale characterized by a set of fuzzy meanings of lexical terms. On the right the fuzzy description of a value x_0 produces a LFS.

$LF(L)$ contains any fuzzy description $D(x_0)$ of a physical state x_0 as shown in fig. 1.

Whereas the only relation that can be defined on symbols (or terms) is the equality relation, more interesting relations can be defined on fuzzy subsets of linguistic terms [6]. We especially point out the transportation distance operator (denoted d'_S) proposed in [7] that allows to compute the distance between 2 LFSs. This distance d'_S on the space of LFSs has the property to be an extension of a distance d_S on the space of linguistic terms. Actually the existence of this distance defines the scale as a fuzzy metrical scale, i.e. a scale that preserves a distance from the measured quantity to its representation:

d_S is a distance defined on terms s of L,

d'_S is a distance defined on $D(X)$ such that the singleton coincidence with d_S is respected: $d'_S(\{a\},\{b\}) = d_S(a,b)$.

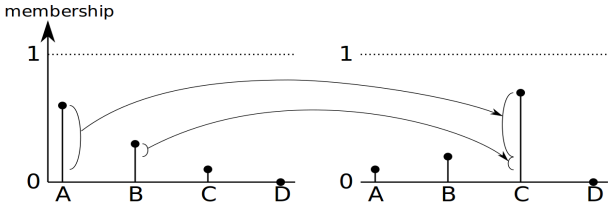

Fig. 2. The transportation distance between the 2 LFSs is defined as the minimal cost of the transportation of membership degrees between terms in order to get the second LFS from the first one.

To get the transportation distance d'_S between 2 LFSs, a cost of membership transportation is computed. Each transportation of an amount a of membership from a term A to a term B is associated to the cost:

$$c(a,A,B) = a.d_S(A,B) \qquad (2)$$

The transportation distance is then defined as the minimal cost to bring membership amounts.

The distance d_S on terms can be experimentally defined by a calibration process, but can also be defined to include some context knowledge. Indeed, the distance between 2 terms considered as quite similar is arbitrary defined as small in order to take this similarity into account.

Another usage of this distance is the definition of a confidence volume playing for the LFSs the same role than the confidence interval for numerical values.

Fuzzy scale gives a tool to represent a gesture with a lexical fuzzy subset (LFS), but gives no semantic to the membership degree of each term to this fuzzy subset. The constraint given by (1) can suggest a Bayesian semantic for these degrees but the building process doesn't involve any Bayesian hypothesis. In this study, we decided to choose a weak semantic by interpretation of fuzzy membership degrees as belief masses as defined in the TBM (Transferable Belief Model) [8]. Each LFS is now termed as Bayesian Basic Belief Assignment because the membership degree of a LFS is interpreted as a mass of 1 distributed on the singletons in L.

B. Human gesture representation

In the field of human system interaction, it is big interest in developing natural interfaces using human communication modalities. One of these modalities is communication through gestures. Gestures can be described with words. Fig. 3 illustrates the gesture "come here". In this example, the "come here" configuration is described quite simply as the one with a straight arm, and arm is middle down and fore. This particularity of gestures makes possible to describe them with fuzzy nominal scales.

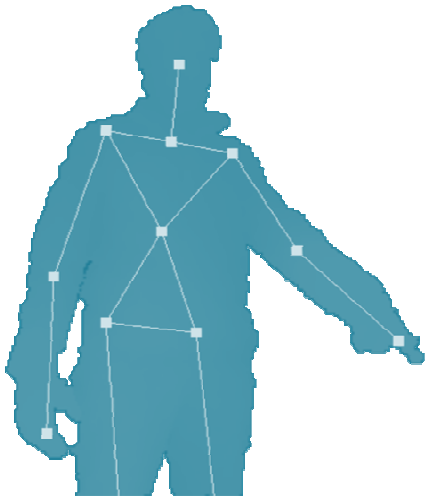

Fig. 3. The "come here" gesture can be expressed with word.

Fig. 4 illustrates a 2D representation of the body perceived by the depth sensor and of the measured joints. Posture detection is produced from the depth camera sensor (Xtion Pro Live from Asus) using NiTE library. The output is a set of 3D joins positions that we call measured skeleton.

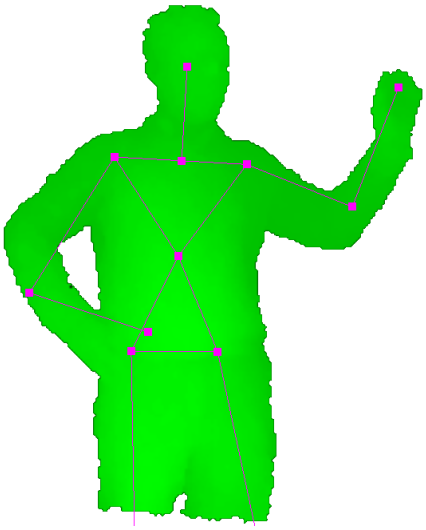

Fig. 4. 2D representation of 3D measurements of joints.

We illustrate our approach with a right upper limb detection gesture. To express the human gesture representation, we decompose the representation into 2 intermediate high level representations corresponding to arm and forearm of measured skeleton.

The angles relative to the arm and the forearm are computed from the joints positions of detected skeleton. From the physical measurement, each body part (i.e. the arm and the forearm) are expressed using word. Lexical sets are chosen to have simple and easily understandable descriptions. Each part (i.e. the arm and the forearm) is characterized by the 2 measured angles issued from joints positions. They are illustrated in fig. 5. We also use lexical fuzzy subsets as illustrated in fig. 1.

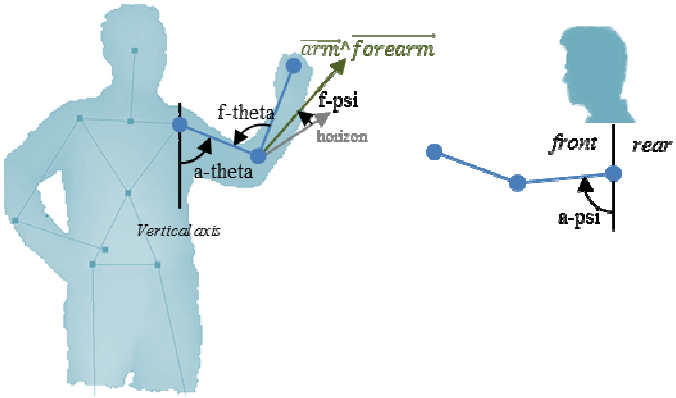

Fig. 5. Used angles from skeleton.

1) Arm modeling

For arm angle measurement, we use *a-tetha* angle and *a-psi* angle. *a-tetha* angle corresponds to the angle formed from the vertical to the arm, in the plane of the human body. *a-psi* is the angle formed from the vertical to the arm when the arm is level up front of the human. Lexical fuzzy subset of *a-theta* is *{down, horizon, up}* and the lexical fuzzy subset of *a-psi* angle is *{rear, middle, front}*. Fig. 6 illustrates the *a-theta* angle lexical posture determination from the angle measurements.

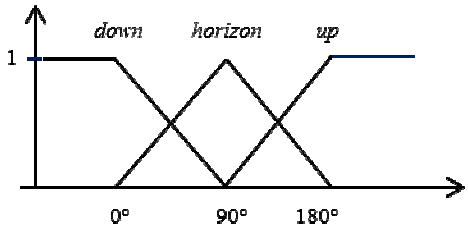

Fig. 6. First sub-lexical for arm posture from *a-theta* angle value

Fig. 7 illustrates the *a-psi* angle lexical posture.

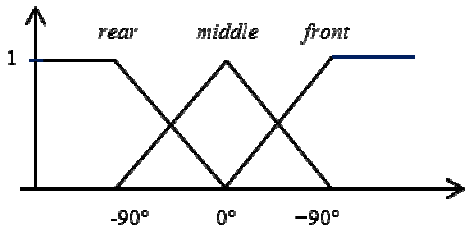

Fig. 7. Second sub-lexical for arm posture from *a-psi* angle value.

Then, the arm posture is represented by the combination of these 2 lexical subsets, corresponding of the 2 angles *a-tehta* and *a-psi.*

The arm corresponding variable takes its values in the lexical set L_{arm}:

$$L_{arm} = \{down, front, up, aside, rear\} \quad (3)$$

Table 1 gives the arm position rules from these lexical subsets.

TABLE I. ARM GESTURE DEFINITION RULES

		a-tetha		
		down	***horizon***	***up***
a-psi	***rear***	down	rear	up
	middle	down	aside	up
	front	down	front	up

2) Forearm modeling

A similar process of arm modeling is done for the forearm part. 2 angles are used: *f-tetha* which corresponds of angle formed by the forearm axis and arm axis and *f-psi* which corresponds to the position relative to the horizon. Fuzzification process is the same as arm part and is not detailed in this paper.

The forearm corresponding variable takes its values in the lexical set $L_{forearm}$:

$$L_{forearm} = \{open, vclose, hclose, vmiddle, hmiddle\} \quad (4)$$

Table 2 represents the position rules from lexical subset corresponding to *f-theta* and *f-psi*.

TABLE II. FOREARM GESTURE DEFINITION RULES

		f-tetha		
		close	***middle***	***open***
f-psi	***vertical***	vclose	vmiddle	open
	horizontal	hclose	hmiddle	open

3) Right upper limb modeling

The gesture is represented by the combination of an arm posture and a forearm posture. Gesture is also defined by words. Each word represents a human gesture.

The lexical set of gesture is defined by a list of modal gestures $L_{G\text{-}modal}$ presented in (5).

Using rules, each modal gesture made by upper limb posture can be defined in the system. This learning step is based on the sub-gesture entries: the arm and forearm values. Table 3 illustrates rules defining gestures to be detected from lexical subset defined in (3) and (4).

It is noted that the example in table 3 illustrates a 2D dimension case due to the *2* linguistic variables (corresponding to arm and forearm). It is possible to generate the principle to *n* dimensions.

TABLE III. MODAL GESTURE DEFINITION RULES

		arm				
		down	***front***	***up***	***aside***	***rear***
forearm	***open***	down	front	up	aside	rear
	vclose	*ND*	ND	ND	Touch Ear	ND
	hclose	ND	ND	ND	ND	ND
	vmiddle	shake	ND	ND	psi	ND
	hmiddle	ND	protect	ND	ride	ND

ND : not defined

Note that the configuration of arm and forearm postures can be not corresponding to any modal gesture (ND in table 3). For this example, the linguistic set of modal gesture is given in (5).

$$L_{G\text{-}modal} = \{down, front, up, aside, rear, psi, ride, touchEar, shake\} \quad (5)$$

The known gestures presented in this paper are modal gestures. Gestures are expressed into linguistic representation. They are obtained from the measurement of human postures. In order to simplify the approach, we use directly the posture detection to define a static gesture. Dynamic gestures are obtained using the temporal dimension of the sensor measurements. Anyway, as soon as the gesture is computed, further modeling and processing are identical. The next section details the posture recognition process.

III. FUZZY POSTURE RECOGNITION PROCESS

To perform the gesture recognition process, the system has to know which gestures can be detected. Thus, the result of detection process of any gesture can be expressed in terms of mass on known gestures [9][10]. The mass criterion is computed from the detected gestures and each known gestures. The linguistic descriptions provided by the gesture recognition process form the linguistic description of this right upper limb posture. Hence, this posture *P* (considered as a gesture) can be written as:

$$P = (L_{G\text{-}modal,1}, L_{G\text{-}modal,2}, L_{G\text{-}modal,3}, \ldots, L_{G\text{-}modal,n}) \quad (6)$$

where *n* is the number modal gestures contained in $L_{G\text{-}modal}$.

Table 4 illustrates the arbitrary gesture detection case. In this example, the expressed gesture representation is

$$P = (P: 0, Q: 0, R: 0.42, S: 0.28, T:0.18, U: 0.12) \quad (7)$$

Where *P* represents the fuzzy representation of the detected posture presented in table 4 and illustrated in fig. 8.

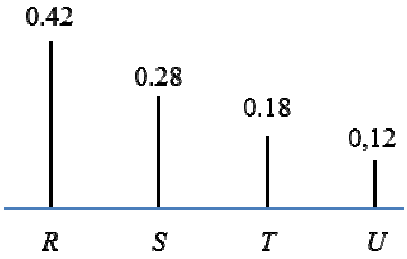

Fig. 8. Selection of action: a basic approach

TABLE IV. ARBITRARY GESTURE DETECTION RESULT

			arm				
			down	***front***	***up***	***aside***	***rear***
			0	**0**	**0.6**	**0.4**	**0**
forearm	***open***	**0**	ND	ND	ND	*P*	*P*
	vclose	**0.7**	ND	*Q*	*R*	*S*	*S*
	hclose	**0.3**	ND	ND	*T*	*U*	ND
	vmiddle	**0**	ND	ND	ND	ND	ND
	hmiddle	0	ND	ND	ND	ND	ND

From the fuzzy gesture detection (fig. 8) process result (represented by (6)), the decision system takes place. Next section presents the decision process.

IV. DECISION SYSTEM UNDER USER CONSTRAINTS

In our application context, the decision process has to choose an action from human gesture detection result. This result is expressed using fuzzy representation. That lets us to express finely the detection: several gestures can be detected from a single human gesture.

This fluency of information carried by fuzzy representation is conventionally employed in the case using multi-sensor detection. In this paper, we focus its interests in order to automatize the choice of action to be done from one human gesture detection. The imprecision carries information that represents the possible confusion, hesitation or belief of detected gestures. The output of decision process stems of defuzzification process.

From one single human gesture detection, several modal gestures are detected, as shown in (7). We consider that experiment context that each gesture corresponds to an action to be executed by the robot NAO. For each modal gesture $L_{G\text{-}modale,i}$ is a corresponding action named A_i.

In a first basic approach, the decision is given by the maximum pignistic probability and then the recognized gesture is the modal gesture which has the most important mass. This is the case of the example 1 in of fig. 9.

In a second approach, we consider the uncertainty given by the measurement process. The main source of uncertainty is issued from the camera process detection. The uncertainty estimation is complex is this case and is not presented in this paper. In any case, the epistemic uncertainty, i.e. the uncertainty relative to gesture definition, is more important and taken into account by the consumer. We consider that the consumer accept a confidence volume around each modal gesture on the space of LFSs. This confidence volume plays for the space of LFSs the same role than the confidence interval for numerical spaces. It is defined by the distance between a measured gesture and the modal gesture. The distance between LFSs is the transportation distance as presented in II.A. It is simply computed when one of the LFSs is a modal gesture. Such distance is then equal to the transportation cost of an amount of mass from the modal term to the other terms. And if all the distances between terms are equal to 1, the distance is given by the amount of mass transported from the modal term to the other ones during the distance computation. The recognized modal gesture is a gesture that has the measured gesture inside its confidence volume. Due to the property of the transportation distance, this is equivalent to the definition of a minimum amount of mass to recognize a gesture. In fig. 9, the case 2 defines 0.6 as a minimum of mass for the recognition and no gesture is recognized. This means that the measured gesture is out of the confidence volumes of each modal gesture. In case 3 the minimum amount of mass is defined as 0.4 then the confidence volume around *R* is large enough to include de measured gesture.

In a third approach, we want to consider that the consumer wants to take into account the possible confusion between 2 possible gestures. The confusion between 2 modal gestures is given by the overlapping of their respective confidence volume. The overlapping depends on the distance between the 2 modal gestures and on the distance that characterizes the confidence volume. So, for a given size of the confidence volume, the overlapping depends only on the distance d'_S between the 2 modal gestures themselves represented by singletons. In accordance with the singleton coincidence of d'_S with d_S , the overlapping depends only on the distance d_S between the 2 modal terms that characterize the 2 modal gestures. As the distance between terms is arbitrary defined, it can be done by the consumer in order to express the need to take care about a possible confusion. As for the previous approach, the corresponding decision is expressed as thresholds. In this approach, a relative threshold synthesizes the possible confusion between 2 modal gestures. In fig. 9 case 4, the *confusion relative threshold* is 0.2 between the terms *R* and *S*. With the measured gesture given in example, the difference between *R* and *S* is smaller than the *confusion relative threshold.* So *R* and *S* are valid at this decision step. In order to perform the decision, another constraint can be added. The nature of the robot action associated to command gesture is able to influence the decision process. To illustrate this consideration, we consider two classes of actions (each action corresponds to a modal gesture): one class contains the "classical" actions and the second the "emergency" actions, corresponding respectively to a "classical" modal gesture and an "emergency" modal gesture. So when an "emergency" nominal gesture is detected even with a lower mass than a classical detected nominal gesture, the selected decision is the corresponding "emergency" action. If the nominal gesture *S* is defined as belonging to the "emergency" class, it is chosen in

the case 4. In case 5, the *confusion relative threshold* is not large enough to select *S* instead of *R*.

In the last approach, the consumer constraint reduces the set of possible recognized gestures to the emergency related one. This situation is similar to the second approach, but with only one gesture. As only one gesture appears in the decision process, the confidence volume can be larger than in other approaches. The case 6 in fig. 9 illustrates this approach.

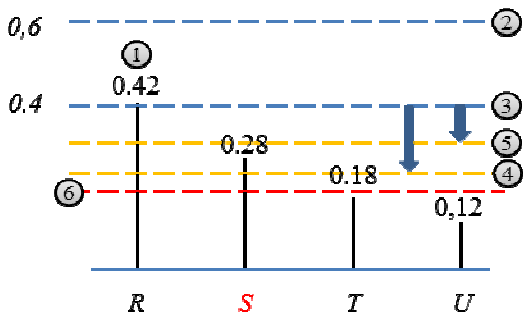

Fig. 9. Examples of threshold effects on decision from an arbitrary gesture detection result.

V. Discussion

The main problem of this approach is to determine the thresholds levels of mass. This step is critical because its directly impact the decision, i.e. action to be done.

To help the user to fix the threshold levels, one way consist to learn it, during a dedicated period of this stage. After this stage, system will be able to select an action, even if the learning process could be continued to improve the determination of threshold in accordance with the user feed-back.

In this section we discuss about facilitating threshold determination based on mass transfers inside fuzzy representation of human gesture.

The symbolic distance between gestures represents their discrimination power. Considering an "emergency" gesture, the distance between the corresponding symbolic gesture to another symbolic gesture, provides information. In the case of a high distance, that means that the cost to switch from a gesture to another one is high. For a low distance, the cost is low. Thus level of transferable mass is higher.

This is especially interesting when open word is adopted, i.e. when unknown gesture is admitted in the fusion system. The unknown gesture is considered to be got close from the "emergency" gesture. To implement this way, a small distance is fixed between these two gestures. Thus the "emergency" gesture is favored.

Finally, the distance between fuzzy representations of gesture gives the transferable mass. Mass corresponding to the unknown gesture is transferred to the emergency gesture in accordance to the distance and the determination of threshold principle is thereby facilitated.

VI. Conclusion and Perspectives

This paper discusses about the importance of human understanding in humans system interaction. An illustrated scenario was implemented in the case of a humanoid NAO robot which is required to make actions by human gesture command. A fuzzy system detection and recognition is constructed to help for gesture recognition more intuitively. In order to automate the decision process, a solution based on the probabilistic and on the evidence theory is presented to take into account the user constraints i.e. rules of decision to be adopted in the case of doubt, hesitancy the priority of actions.

The future scenario of the work will consider several gesture recognition sensors and several kind of sensor. This way illustrates the interest of multi-modal fusion approach which is able to take into consideration conflicting information. In addition, we will consider the way to make a human gesture: the temporal repetition of a gesture or the execution speed of it. That is expressed another kind of user constraints.

References

[1] S. Mitra, S., T. Acharya, "Gesture Recognition: A Survey," IEEE Trans. Sys., Man, and Cybernetics, Part C: Applications and Reviews, vol. 37, no. 3, May 2007, pp. 311-324.

[2] D. Dubois and H. Prade, "Representation and combination of uncertainty with belief functions and possibility measures". Computational Intelligence, vol. 4, no. 3, pp. 244-264, 1988.

[3] E. Benoit, L. Foulloy, "The role of fuzzy scales in measurement theory", Measurement, vol. 46, no. 8, 2013, pp. 2921-2926.

[4] G. Mauris, E. Benoit, L. Foulloy, "Fuzzy symbolic sensors - from concept to applications", Measurement, vol. 12, 1994, pp. 357–384.

[5] E. Benoit, L. Foulloy, "Fuzzy nominal scales", IMEKO TC7 Symposium, Cracow, Poland, 2002, pp. 21–25.

[6] B. De Baets, R. Mesiar, "Metrics and T-equalities", Journal of Mathematical Analysis and Applications, 267, pp. 531–547, 2002.

[7] T. Allevard, E. Benoit, L. Foulloy, "The transportation distance for fuzzy descriptions of measurement". Metrology and Meas. Syst., vol. XIV no. 1, 2007, pp. 25-37.

[8] Smets, P., "Data fusion in the transferable belief model". Inf. Fusion. Proc. IEEE Int. Conf. on Fusion, vol. , 2000, pp. PS21-PS33.

[9] E. Benoit, S. Perrin, D. Coquin, Body posture measurement in a context of example-based teaching, in J. Phys.: Conf. Ser. Vol. 588, 012055, 2015 [Joint IMEKO TC1-TC7-TC13 Symp., Madeira, Portugal, 2014].

[10] T. Allevard, E. Benoit, L. Foulloy, "Dynamic gesture recognition using signal processing based on fuzzy nominal scales", Measurement, vol. 38, no. 4, 2005, pp. 303-312.