Data Fusion in Multimodal Interface

Alexander Alfimtsev
Department of Information Systems and Telecommunications, BMSTU, 2nd Baumanskaya St., Moscow, Russia

Keywords: Multimodal Interface, Data Fusion, Fuzzy Aggregation.

Abstract: A method of data fusion in multimodal interface using fuzzy fusion operators is described in this paper. Special scenes are determined by dynamic and static patterns for which the sources of information are the results of pattern recognition by low-level algorithms.

1 INTRODUCTION

In general case the recognition tasks in multimodal interfaces are not limited by only one specific pattern each instant. At the moment the research is widely held in the field of so-called multimodal recognition. A modality is usually meant as a human form of influence on another human being or a personal computer using speech, gestures, touch, mimics, appearance etc. It is considered now (Averkin, 1986); (Higuchi, 2004); (Sharma, 2003) that even within the bounds of only one interaction form (with the PC), for example, the control of an interface via hand gestures, different modalities can be used. In this case there appears a task of combining, or as often used, fusion of different modalities. The intention to combine several sources of information can be explained by the fact that each source separately may have high uncertainty or inaccuracy of data, which are decreased by their fusion (Wan04). The goals of multimodal recognition may be different. One of them is the control of physical objects based on scenes analysis which are actually relations of dynamic and static objects.

The fusion can be performed at two levels (Gra00): low level and high level. Let's suppose that each signal, i.e. a sequence of n+1 counts $Y_i[t_0, t_n]=\{y_i(t_0), y_i(t_1), y_i(t_2), ..., y_i(t_n)\}$ is connected with its modality. The fusion that deals with signals is usually referred to the low level. Signals and modalities that correspond to them are synchronized on the low level, the interconnection and interaction of signals can be clearly seen, modalities are often referred to the same interaction form. The fusion of the high level is done usually after the work of recognition algorithms on the low level. Each of them realizes the recognition of the group of signals referred to the same form of interaction or even to the same modality. Forms of modalities can be independent of time.

Usually functions used for fusion are called fusion operators (Grabisch and Roubens, 2000). Max-operator is one of the most well-known fusion operators. The use of max-operator for multimodal recognition provides with a high level of reliability but can be ineffective if used on the high level.

Weighted arithmetic operator is another popular fusion operator. But fusion using weighted arithmetic operator may lead to insufficient recognition accuracy where recognition is understood as the percent of successful recognitions from the total number of attempts. This can be a consequence of the empirical choice of weighting coefficients and also the difficulties for considering a possible interconnection of membership functions that characterize the pattern recognition result.

The method that uses fuzzy fusion operators (Sugeno and Choquet fuzzy integral) for data fusion and multimodal recognition of video scenes defined by dynamic and static patterns for which recognition results by the low-level algorithms are the sources of information (secondary attributes (Devyatkov and Alfimtsev, 2008)) is described.

Sugeno and Choquet fuzzy fusion operators are considered in Section 2. The procedures required for data fusion based on fusion operators are described in Section 3. Multimodal scene analysis is considered in Section 4.

Alfimtsev A..
Data Fusion in Multimodal Interface.
DOI: 10.5220/0004365903060310
In *Proceedings of the International Conference on Biomedical Electronics and Devices* (MHGInterf-2013), pages 306-310
ISBN: 978-989-8565-34-1

2 FUZZY FUSION OPERATORS

Fuzzy fusion operators that allow us to consider the interconnection of membership functions use a *fuzzy measure*. The fuzzy measure is a function $g: 2^R \to [0,1]$, where R is a set of some parameters that characterize some object. The fuzzy measure $g(Q_i)$ characterizes a total significance of parameters that are included in the set Q_i. The fuzzy measure satisfies the set of conditions (Ave86): specifically $g(\varnothing) = 0,\ g(Y) = 1;$ if $Q, P \in Y$ and $Q \subset P$ then $g(Q) \le g(P)$.

If R is a set of all subsets of modalities subset $Y = \{Y_1, ..., Y_m\}$ then fusion operators can be in the following way.

Fuzzy Sugeno operator:

$$A_k = A_k^C = \max_{i=1}^{i=m} [\min(\mu_i^k, g(Q_i)] \quad (1)$$

where

$$\mu_1^k(y_1) \ge \mu_2^k(y_2) \ge ... \mu_m^k(y_m),\ Q_i = \{Y_1, ..., Y_i\},\ i = 1, ..., m$$

Fuzzy Choquet operator:

$$A_k = A_k^{III} = \sum_{i=1}^{i=m} [\mu_i^k(y_i) - \mu_{i+1}^k(y_{i+1})] g(Q_i) \quad (2)$$

where

$$\mu_1^k(y_1) \ge \mu_2^k(y_2) \ge ... \mu_m^k(y_m),\ Q_i = \{Y_1, ..., Y_i\},\ i = 1, ..., m,\ \mu_{m+1}^k(y_{m+1}) = 0$$

Fuzzy Choquet operator is usually interpreted as a generalization of the weighted arithmetic average notion and Sugeno operator as generalization of the weighted median concept (with fusion of no less than three modalities).

The most popular due to their simplicity are the methods for calculation of the fuzzy measure based on the concept of g_λ-fuzzy measure introduced by Sugeno. Fuzzy measure is called g_λ-fuzzy measure if the following condition is true for it: for all $Q, P \subset Y$ that $Q \cap P = \varnothing$ occurs $g(Q \cup P) = g(Q) + g(P) + \lambda g(Q) g(P)$ for some $\lambda > -1$.

Let's consider the procedure of the most popular method for g_λ-fuzzy measure (Averkin, 1986); (Devyatkov and Alfimtsev, 2008): (Grabisch and Roubens, 200); (Marichal, 2000) calculation still labeling it as g.

Step 1. For each signal (modality) Y_i, $i = 1, ..., m$ select the value of fuzzy measure $g(Y_i) \in [0,1]$ as an importance degree of the modality Y_i. Values $g(Y_i)$ can be set by an expert, can be a result of an experiment or can be received another way.

Step 2. Find a value λ using the equitation (3).

$$\lambda + 1 = \prod_{i=1}^{m} (1 + \lambda g(Y_i)) \quad (3)$$

Step 3. For all $Q_i = \{Y_1, ..., Y_i\},\ i = 1, ..., m$ find recursive fuzzy measures $g(Q_i)$ using the following expressions:

$$g(Q_1) = g(Y_1),$$
$$g(Q_i) = g(Y_i) + g(Q_{i-1}) + \lambda g(Y_i) g(Q_{i-1}), \quad (4)$$
$$i = 2, ..., m.$$

3 MULTIMODAL RECOGNITION

Before considering a common procedure of data fusion let's formalize the procedure of composition of the set Y_i and the procedure of recognition using the algorithm i with function $\mu(y_{ij_i})$. In the general case the sources for fusion are i algorithms, i = 1,…,m that use hidden modalities. In this case hidden modalities and the methods for their fusion are not considered. The things that are of interest - the results of the work of each algorithm as a source of a new separate signal (modality) $Y_i, i = 0, ..., m$ and a membership function $\mu(y_{ij_i}), y_{ij_i} \in Y_i, i = 0, ..., m,\ j_i = 0, ..., n_i$.

The main task of the procedure is fusion of modalities $Y_i, i = 0, ..., m$. Each algorithm passes a preprocessing according to the next procedure 1 in

order to form a set Y_i and membership functions $\mu(y_{ij_i}), y_{ij_i} \in Y_i, i = 0,...,m,\ j_i = 0,...,n_i$.

Step 1. A combination of empty sets $Y_i^k = \varnothing, k = 1,...,K$ is specified.

Step 2. For each reference object *k*, *k*=1,…, *K*, a reference model $G_i^k, k = 1,...,K$ is formed using the hidden modalities.

Step 3. Model *G* is formed for the recognizable object based on the same principles and modalities.

Step 4. Model *G* compared to each model $G_i^k, k = 1,...,K$ resulting the calculation of the set of counts $\{y_i^1, y_i^2,..., y_i^K\}$ which characterize the proximity of model *G* to models $G_i^k, k = 1,...,K$ respectively.

Step 5. Sets $Y_i^k \cup y_i^k$, $k = 1,...,K$ are formed which are considered as the new sets Y_i^k. If sets Y_i^k stop to change then go to step 6 (other criterions can be used to go to step 6). Else the procedure is started from step 2.

Step 6. Sets Y_i^k are joined resulting the set of $Y_i = \bigcup_{k=1}^{K} Y_i^k$ which is sorted (if it's numeral then the ascending sort is done) and its elements are indexed $i = 1,...,m,\ j_i = 0,...,n_i$ resulting a set $Y_i = \{ y_{ij_i} \in Y_i \,|\, i = 1,...,m,\ j_i = 0,...,n_i \}$. A membership function $\mu(y_{ij}), y_{ij} \in Y_i, i = 1,...,m,\ j = 0,...,n_i$ is specified on the set Y_i.

Recognition based on the separate algorithm *i* with function $\mu(y_{ij_i})$ can be done according to the following procedure 2.

Step 0. The forming of set Y_i and membership function $\mu(y_{ij_i}), y_{ij_i} \in Y_i, i = 0,...,m,\ j_i = 0,...,n_i$ is done with procedure 1.

Step 1. For each reference object *k*, *k*=1,…, *K* their own reference model is done using hidden modalities.

Step 2. Model G is formed for the recognizable object based on the same principles and modalities.

Step 3. Model G is compared with each model $G_i^k, k = 1,...,K$ resulting a calculation of the set of counts $\{y_i^1, y_i^2,..., y_i^K\} \subset Y_i$ which characterize the proximity of model G to models $G_i^k, k = 1,...,K$ respectively.

Step 4. Model *G* is considered concurrent with that reference model G_i^k for each the value $\mu(y_i^k),\ y_i^k \in Y_i$ is maximum.

Thus the membership function $\mu(y_i^k),\ y_i^k \in Y_i$ estimates the proximity of the recognizable model to the corresponding reference model. The main task is fusion of modalities $Y_i, i = 1,...,m$ to increase the reliability of recognition.

Thus the common method for data fusion in multimodal interface using Sugeno and Choquet operators will be the following.

Step 1. For each modality (signal) $Y_i, i = 1,...,m$ check the value $g(Y_i) \in [0,1]$ as an importance degree of modality Y_i.

Step 2. Find value λ using equitation (3).

Step 3. Calculate a set of membership functions $\mu(y_i^k),\ y_i^k \in Y_i,\ i = 1,...,m.$ using procedure 2 for the recognizable object for each algorithm $i = 1,...,m$ and for each $k = 1,...,K$.

Step 4. For each $k = 1,...,K$ sort a set of functions $\mu(y_i^k)$ so $\mu(y_{j_1}^k) \geq \mu(y_{j_2}^k) \geq ... \geq \mu(y_{j_m}^k), j_n \in \{1,...,m\}$.

Step 5. For each $k = 1,...,K$ calculate fuzzy measures values $g(Q_i^k)$ recursively, where $Q_i^k = \{Y_{j_1},...,\ Y_{j_i}\},\ i = 1,...,m$ using equitation (4).

Step 6. Calculate operator values $A_k = A_k^C$ (or $A_k = A_k^{\text{III}}$) for all $k = 1,...,K$. The recognizable object is considered concurrent with the reference object for which the value $A_k = A_k^C$ is maximum.

4 RECOGNITION OF VIDEO SCENES

We hope you find the information in this template useful in the preparation of your submission.

In a general case each frame can contain L objects $\theta_l,\ l=1,...,L$ which are subject to recognition. Objects from different sets Θ_l can be in defined, in a general case r-nary relations $\Xi \in \Theta_{l_1} \times \Theta_{l_2} \times ... \times \Theta_{l_r},\ \{l_1,\ l_2,...,\ l_r\} \subseteq \{1,...,\ L\}$ Each of these relations Ξ we'll call the reference scene. By analogy with the recognizable object we'll identify the recognizable scene $\xi = <\theta_{l_1},\ \theta_{l_2},... ,\theta_{l_r}>,\ \{l_1,\ l_2,...,\ l_r\} \subseteq \{1,...,\ L\}$, where $\theta_{l_1},\ \theta_{l_2},...,\theta_{l_r}$ are the recognizable objects and the process of recognizing similarities with some reference scene Ξ we'll call the recognition of the scene ξ .The similarity of the recognized scene ξ with the reference scene Ξ characterized with the nonzero value of the similarity criterion we'll write as $\xi \approx \Xi$. In case the value of the similarity criterion is equal to zero then scene ξ is not similar to scene Ξ. This non-similarity is written this way $\xi \neq \Xi$.

Operator $A^j[\mu_1(y_1),\ \mu_2(y_2),...,\ \mu_m(y_m)]$ is used for calculation of the similarity of the recognizable θ and the reference object Θ^j. Membership functions $\mu_1(y_1),\ \mu_2(y_2),...,\ \mu_m(y_m)$ with values in the interval [0,1] are the arguments of this operator. Operator values also lie in the interval [0,1]. Thus the operator is a function $[0,1]^m \to [0,1]$. Then if it is known for each object of the recognizable scene $\xi = <\theta_{l_1},\ \theta_{l_2},... ,\theta_{l_r}>$ a set of criterion values $A_{l_1},\ A_{l_2},...,\ A_{l_r}$ of its similarity with the objects $\Theta_{l_1}^{k_{l_1}},\ \Theta_{l_2}^{k_{l_2}},...,\ \Theta_{l_r}^{k_{l_r}}$ of the reference scene then using some fusion operator A we can calculate the similarity measure between the recognizable scene and the reference scene as a value of function $A[A_{l_1},\ A_{l_2},...,\ A_{l_r}]$.

Procedure 4 of recognition of the separate scene (relation) $\xi = <\theta_{l_1},\ \theta_{l_2},... ,\theta_{l_r}>,\ \{l_1,\ l_2,...,\ l_r\} \subseteq \{1,...,\ L\}$ that uses this idea will look as follows.

Step 1. Each object $\theta_{l_1}, \theta_{l_2},...,\theta_{l_r}$ is recognized separately comparing with reference objects $\Theta_{l_1}^{k_{l_1}},\ \Theta_{l_2}^{k_{l_2}},...,\ \Theta_{l_r}^{k_{l_r}}$, $k_{l_r} = 1,...,\ K_{l_r}$, $\{l_1,\ l_2,...,\ l_r\} \subseteq \{1,...,\ L\}$ using fusion operators $A_{l_1},\ A_{l_2},...,\ A_{l_r}$. If for all recognizable objects $\theta_{l_1}, \theta_{l_2},...,\theta_{l_r}$ similar reference objects $\Theta_{l_1}^{\tilde{k}_{l_1}},\ \Theta_{l_2}^{\tilde{k}_{l_2}},...,\ \Theta_{l_r}^{\tilde{k}_{l_r}}$ $\tilde{k}_{l_r} = 1,...,\ K_{l_r}$, with them were found such as $\theta_{l_1} \approx \Theta_{l_1}^{\tilde{k}_{l_1}},\ \theta_{l_2} \approx \Theta_{l_2}^{\tilde{k}_{l_2}},...,\ \theta_{l_r} \approx \Theta_{l_r}^{\tilde{k}_{l_r}}$ then go to step 2. If there was found no similar reference object for at least one object $\theta_{l_1}, \theta_{l_2},...,\theta_{l_r}$ then go to step 3.

Step 2. Scene $\xi = <\theta_{l_1},\ \theta_{l_2},... ,\theta_{l_r}>$ is considered recognized and similar with scene $\Xi = \Theta_{l_1}^{\tilde{k}_{l_1}},\ \Theta_{l_2}^{\tilde{k}_{l_2}},...,\ \Theta_{l_r}^{\tilde{k}_{l_r}}$ and the value of the criterion for scene similarity is equal to $A[A_{l_1}^{\tilde{k}_{l_1}},\ A_{l_2}^{\tilde{k}_{l_2}},...,\ A_{l_r}^{\tilde{k}_{l_r}}]$.

Step 3. Scene $\xi = <\theta_{l_1},\ \theta_{l_2},... ,\theta_{l_r}>$ was not recognized.

We'll call scenes

$\Xi \in \Theta_{l_1} \times \Theta_{l_2} \times ... \times \Theta_{l_r},\ \{l_1,\ l_2,...,\ l_r\} \subseteq \{1,...,\ L\}$

the scenes of the 1st level and identify them Ξ_1. We'll call scenes $\Xi_s \in \Xi_{s-1}^1 \times \Xi_{s-1}^2 \times ... \times \Xi_{s-1}^v$ as scenes of the s-layer, where $\Xi_{s-1}^1, \Xi_{s-1}^2,... ,\Xi_{s-1}^v$ - scenes of the (s-1) layer. Thus the scenes of the 1st layer are the relations of the objects and scenes of s-layer, where s>1 are relations of the scenes of (s-1)-layer. In order to recognize (s-j)-level scenes (j=0,1,…, s-2) it is needed to recognize the scenes of (s-j-1)-level the relation of which are (s-j)-level scenes. If during the recognition of any (s-j)-level scene it is found that at least one (s-j-1)-level scene included in the relation of this (s-j)-level scene can't be recognized then the recognition process of the latter is stopped.

The development of the procedure 4 of the

recognition of the 1st-level scenes can be put in the base of the method of s-layer scenes $\Xi_s \in \Xi_{s-1}^1 \times \Xi_{s-1}^2 \times \ldots \times \Xi_{s-1}^v$ recognition as follows.

Step 1. Each object $\theta_{l_1}, \theta_{l_2}, \ldots, \theta_{l_r}$ that is included at least in one 1st-level scene Ξ_1 is recognized by separate comparison with reference objects $\Theta_{l_1}^{k_{l1}}$, $\Theta_{l_2}^{k_{l2}}, \ldots, \Theta_{l_r}^{k_{lr}}$ $k_{l_r} = 1, \ldots, K_{l_r}$, $\{l_1, l_2, \ldots, l_r\} \subseteq \{1, \ldots, L\}$ using fusion operators A_{l_1}, $A_{l_2}, \ldots, A_{l_r}$.

Step 2. Each 1st-level scene Ξ_1 for all objects of which are found similar reference objects is considered recognized and a similarity criterion is extracted for it (the value of the fusion operator) A_1. After this go to step 3. If there was found no such scenes then there's no recognized scenes of the first level and higher and the execution is stopped.

Step 3. The value of level is set to 2 and we go to step 4.

Step 4. If there were found s-level scenes Ξ_s for all (s-1)-level scenes of which had been found nonzero values of the similarity criterion then these scenes Ξ_s are considered recognized and similarity criteria (fusion operator values) A_s are calculated for them. If there are any (s+1)-level scenes then step 4 is executed once again with the value s=s+1. Else the execution is stopped.

If there were found no s-level scenes Ξ_s for all (s-1)-level scenes of which had been found nonzero values of the similarity criterion then there are no recognized s-level scenes and the execution is stopped.

5 CONCLUSIONS

There is considered a developed method that uses fuzzy fusion operators with fuzzy measure for data fusion in multimodal interface. The main advantages of the method from the well-known analogs are the following:

- The ability of hierarchic multimodal recognition of scenes that consist of static and dynamic (moving) objects.
- The ability to consider the measure of importance of each modality during the process of hierarchic scene recognition due to the use of fusion operators that use a fuzzy measure.
- The ability to be the base for developing of control systems for different objects with the help of dynamic patterns (robots, computers, TV-sets etc.).
- The ability to increase the reliability of recognition of separate objects (for example, a human being) in the scene using relations between these objects and other objects of the scene (background objects).
- Promising opportunities for development of intellectual and intuitive human-machine interfaces by using more modalities and relations.

REFERENCES

Alatan, A. A. Automatic multimodal dialogue scene indexing // *Proc. of image processing*.- 2001.- Vol. 3.- P. 374-377.

Averkin, A. A., Batyrshin, I.Z., Blishun, A.F. Fuzzy sets in control models and artificial intelligence systems // Moscow . *Book, Nauka*, 1986.

Devyatkov V., Alfimtsev A. Optimal Fuzzy Aggregation of Secondary Attributes in Recognition Problems // *Proc. of 16-th Int. Conf. in Central Europe on Comp. Graphics, Visual. and Computer Vision*.-Plzen, 2008.- P. 78–85.

Grabisch M., Roubens M. Application of the Choquet Integral in Multicriteria Decision Making. *In Fuzzy Measures and Integrals*- Theory and Applications, Physica Verlag, 2000, pp. 415-434.

Higuchi M. and oth. Scene Recognition Based on Relationship between Human Actions and Objects // *17th Int. Conf. on Pattern Recog*.-2004.- Vol. 3.- P. 73-78.

Liu F., Lin X. Multimodal face tracking using Bayesian network // *IEEE Internat. Workshop on Analysis and Modeling of Faces and Gestures*.- Nice, 2003.-P. 135-142.

Marichal J. On Choquet and Sugeno Integrals as Aggregation Functions // *In Fuzzy Measures and Integrals*.-2000.-Vol. 40.-P. 247-272.

Ronshin A. L., Karpov A.A., Li I.V. Speech and multimodal interface // Moscow. *Book, Nauka*, 2006.

Sharma R. Speech-Gesture Driven Multimodal Interfaces for Crisis Management// *The IEEE Proc*.-2003.-Vol. 91, №9.- P. 1327–1354.

Wang X., Chen J. Multiple Neural Networks fusion model based on Choquet fuzzy integral // *Proc. of the Third Intern. Conf. on Mach. Learn. and Cybern*.- Vol.4.-P. 2024-2027.