A Priori Data and A Posteriori Decision Fusions for Human Action Recognition

Julien Cumin, Grégoire Lefebvre

A Priori Data and A Posteriori Decision Fusions for Human Action Recognition

Julien Cumin and Grégoire Lefebvre
Orange Labs, R&D, Meylan, France
julien1.cumin@orange.com, gregoire.lefebvre@orange.com

Keywords: Action Recognition, Decision Fusion, Voting Methods, Dempster-Shafer Theory, Possibility Theory.

Abstract: In this paper, we tackle the challenge of human action recognition using multiple data sources by mixing *a priori* data fusion and *a posteriori* decision fusion. Our strategy applied from 3 main classifiers (Dynamic Time Warping, Multi-Layer Perceptron and Siamese Neural Network) using several decision fusion methods (Voting, Stacking, Dempster-Shafer Theory and Possibility Theory) on two databases (MHAD (Ofli et al., 2013) and ChAirGest (Ruffieux et al., 2013)) outperforms state-of-the-art results with respectively $99.85\% \pm 0.53$ and $96.40\% \pm 3.37$ of best average correct classification when evaluating a leave-one-subject-out protocol.

1 INTRODUCTION

In the last decades, human action recognition based on inertial or visual data sources has been an active area of research due to its success in robotics, video games, surveillance, *etc.* Nevertheless, some challenging difficulties still exist, caused by what is called the "3V" issues (IBM et al., 2011): the action *Velocity*, the action *Variety* and the action *Volume*.

Action recognition requires instantaneous responses from the system, moreover if it is an interactive application where actions are used as interaction controllers. In this regard, the velocity of data generation is a major problem, imposing constraints of execution times on algorithms.

Moreover, there is great variability about the way people produce actions. Dynamic variations occur when people produce intense or phlegmatic, slow or fast gestures. Different shapes, orientations and directions may then be captured from body trajectories. These variations exist between people but also for a single user producing the same set of actions (e.g. human ability, left or right-handed, on the move, in different use contexts, *etc.*).

Finally, volumetric variations are challenging as well, ranging from one user in a close world paradigm to multi-users in an open world paradigm.

Consequently, when designing a pattern recognition system, several steps are needed to deal with these issues: processing input data in order to reduce noise and to enhance salient information, clustering data to reduce the dimensionality of the problem, and learning a specific action classifier.

In this paper, we propose a human action classification system mixing *a priori* data fusion and *a posteriori* decision-level fusion (*i.e.* classifier fusion) methods.

This paper is organized as follows. In Section 2, we present some main literature methods on action recognition and decision fusion. In Section 3, we explain in details our fusion strategy. Then, Section 4 shows our experimental configurations and results. Finally, our conclusions are drawn and perspectives are presented in the last section.

2 STATE OF THE ART

2.1 Action Recognition

Human action recognition has been deeply studied for the past ten years. Some studies are based on inertial sensors, others on visual skeleton acquisitions, or sometimes both simultaneously.

Based on inertial data, three main strategies can be identified. The first action recognition strategy relies on similarity metrics between unknown actions to be classified and class reference instances. One main representative (Akl and Valaee, 2010) is a model constructed from the Dynamic Time Warping (DTW) similarity distance and a K Nearest Neighbor (KNN) classifier. Others studies by (Berlemont et al., 2015) proposed a non-linear metric learning strategy based

on Siamese Neural Networks (SNN). The second strategy consists in a statistic modeling approach with Hidden Markov Models (HMM), as in (Pylvänäinen, 2005) in order to model correlations between temporal data samples. Finally, the last strategy implies machine learning methods in order to model class features, such as Support Vector Machines (SVM) (Wu et al., 2009), Bayesian Networks (Cho et al., 2006) or Recurrent Neural Networks (Lefebvre et al., 2015).

Using visual feature data, these three main strategies are still relevant. Firstly, (Zhou and De la Torre Frade, 2012) present a Generalized Time Warping (GTW) algorithm, which is an extension of the DTW algorithm to temporally align multi-modal sequences from multiple subjects performing similar activities. Secondly, (Xia et al., 2012) present an approach for human action recognition with histograms of 3D joints locations. These features are projected using Linear Discriminant Analysis (LDA) and clustered into several posture visual words. The temporal evolutions of those visual words are then modeled by a discrete HMM. Thirdly, a study by (Vemulapalli et al., 2014) uses a SVM classifier to build an action recognition system. Their approach is based on a skeletal representation modeling the 3D geometric relationships between body parts using rotations and translations in 3D space. Since 3D rigid body motions are members of the Special Euclidean group SE(3), human actions can be modeled as curves in this Lie group.

These previous strategies focus on one sensor and one classifier to increase action recognition rates. This can be viewed as a classifier selection. Few studies take into account multi-modal sources and several classifiers to build a more robust system. (Chen et al., 2015) present a two-level fusion approach based on two modality sensors consisting of a depth camera and an inertial body sensor. In the feature-level fusion, features generated from the two differing modality sensors are merged before classification. In the decision-level fusion, outcomes from two classifiers are combined with decision fusion (in their article, using Dempster-Shafer Theory (DST)).

Inspired by this recent study, we propose to fuse *a posteriori* decisions taken by classifiers, on the one hand from individual data, and on the other hand from *a priori* combined data.

2.2 Decision Fusion

In the following (see Figure 1), we assume that the final classification should be made between c_i classes, with $i \in \{1,\dots,I\}$. We have available C_j classifiers, with $j \in \{1,\dots,J\}$, each giving a decision $x^j_{i,k} \in [0,1]$ for a class c_i about a gesture instance $G_k, k \in \{1,\dots,K\}$. A decision $x^j_{i,k}$ closer to 1 indicates a high confidence that the instance belongs to the class c_i, whereas a decision closer to 0 indicates a low confidence. The final decision taken by the decision fusion method is denoted as c_d.

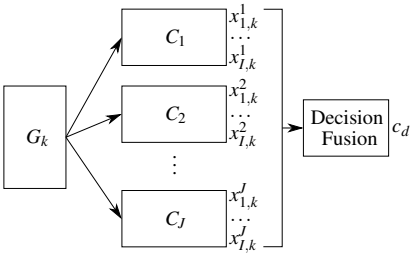

Figure 1: The decision fusion process.

2.2.1 Voting Methods

Voting methods are based on the following principle: each classifier C_j adds a vote V^j_i for each class c_i. The class decided by the fuser is then the one that collects the most votes. In case of a tie, a class at random from those with the most votes is chosen.

Voting by Majority (VM) C_j adds a vote of 1 for the class it has the most confidence in (*i.e.* the class c_i for which $x^j_{i,k}$ is closest to 1), and a vote of 0 for all other classes.

Voting by *Borda Count* (VBC) C_j adds a vote of $P \in \{1,\dots,I\}$ for the class it has the most confidence in, $P-1$ for the second most confident class, ..., 1 for the P^{th} most confident class, and 0 for all remaining classes for which it has less confidence. If $P = 1$, this VBC method is identical to VM.

Weighted Votes (VW) C_j adds a vote of $V^j_i = \varphi(x^j_{i,k})$, with φ a weighting function, taking into account the decision value. As in the *Borda Count* voting method, only the $P \in \{1,\dots,I\}$ classes the classifier has the most confidence in can vote, while the remaining classes receive a vote of 0.

Voting by *Kumar and Raj* (VK) A weighted vote approach is also presented by (Kumar and Raj, 2015). We define the positive set X^+_i (*i.e.* instances belonging to a class c_i), the negative set X^-_i (*i.e.* instances belonging to all the other classes) and $\beta \in \mathbb{R}$ a regularization parameter. $x^j_{i,k}$ is set to 0 if c_i is not the best decided class for C_j (meaning only the best class for

each classifier is voting) and the total number of votes V_i of a class c_i is then:

$$V_i = \sum_{i=1}^{I} w_{i,k}^j x_{i,k}^j, \text{ where} \tag{1}$$

$$w_{i,k}^j = \underset{w}{\text{argmax}} \ |R(x_{i,k}^j, w)|, \text{ with} \tag{2}$$

$$R(x_{i,k}^j, w) = \frac{1}{|X_i^-|} \sum_{u \in X_i^-} \frac{1}{1+e^{-\beta w^\intercal (x_{i,u}^j - x_{i,k}^j)}} - \frac{1}{|X_i^+|} \sum_{u \in X_i^+} \frac{1}{1+e^{-\beta w^\intercal (x_{i,k}^j - x_{i,u}^j)}}. \tag{3}$$

2.2.2 Stacking Methods

Stacking methods (Wolpert, 1992) are based around the fact that decision fusion is equivalent to a classification task: a fusion classifier has to correctly classify an instance using the decisions of initial classifiers as inputs. We then build a feature vector to be learnt by stacking methods as the concatenation of the first-level decisions. The fusion classifier can be trained on the same data set that was used to train the initial classifiers. A multitude of different classifiers can be used to perform stacking. We use two in this paper: MultiLayer Perceptron and Support Vector Machine, respectively referenced by SMLP and SSVM in the experimental section (see section 4).

2.2.3 Dempster-Shafer Theory (DST)

Dempster-Shafer Theory, also called *Evidence Theory*, models data uncertainty and imprecision with what is called *mass functions*.

Let $\Theta = \{c_1, c_2, \ldots, c_n\}$ be the set of classes of the problem, and let $2^\Theta = \{\emptyset, \{c_1\}, \{c_2\}, \{c_1, c_2\}, \{c_3\}, \{c_1, c_3\}, \ldots, \Theta\}$ be the power set of Θ. We define a *mass function* m_j associated to a source C_j, as follows:

$$m_j : 2^\Theta \longmapsto [0,1], \sum_{A \in 2^\Theta} m_j(A) = 1. \tag{4}$$

In practice here, we use :

$$m_j(A) = \begin{cases} \frac{x_{i,k}^j}{\Sigma_{j=1}^J x_{i,k}^j} & \text{if } A = \{c_i\} \\ 0 & \text{otherwise} \end{cases}. \tag{5}$$

We can combine the mass functions of all classifiers into a single mass function m. Using Smets' combination rule (Smets, 1990), we have:

$$\forall A \in 2^\Theta, m(A) = \sum_{B_1 \cap \ldots \cap B_k = A} \prod_{j=1}^{J} m_j(B_j). \tag{6}$$

The decided class c_d is then the one that maximizes the *plausibility* function Pl:

$$c_d = \underset{c_i, i \in \{1,\ldots,I\}}{\text{argmax}} \ \text{Pl}(\{c_i\}) \tag{7}$$

$$= \underset{c_i, i \in \{1,\ldots,I\}}{\text{argmax}} \left(\sum_{B \in 2^\Theta, B \cap \{c_i\} \neq \emptyset} m(B) \right). \tag{8}$$

2.2.4 Possibility Theory (PT)

Possibility Theory, like DST, is aimed at modeling uncertainty and imprecision in data, based on the theory of fuzzy sets.

Let $\pi_k^j = \{\mu_{1,k}^j, \ldots, \mu_{I,k}^j\}$ be the set of membership degrees of an instance G_k to classes c_i with classifier C_j. The decided class c_d is here the one with the maximum merged membership degree $\mu_{d,k}$, as in Equation 9.

$$c_d = \underset{i \in \{1,\ldots,I\}}{\text{argmax}}(\mu_{i,k}). \tag{9}$$

As in (Fauvel et al., 2007), $\mu_{i,k}$ can be evaluated as:

$$\mu_{i,k} = \sqrt{\frac{\sum_{j=1}^{J} (w^j \mu_{i,k}^j)^2}{J}}, \text{ where} \tag{10}$$

$$w^j = \frac{\sum_{p=0, p \neq j}^{J} H_{0,5}(\pi_k^p)}{(J-1) \sum_{p=0}^{J} H_{0,5}(\pi_k^p)}, \text{ and where} \tag{11}$$

$$H_\alpha(\pi_k^j) = \frac{1}{2^{-2\alpha} I} \sum_{i=1}^{I} (\mu_{i,k}^j)^\alpha (1 - \mu_{i,k}^j)^\alpha. \tag{12}$$

3 OUR STRATEGY

Suppose we have 3 data sources A, B and C to record an instance we want to classify (in our case, a human action), and suppose we have 2 classifiers C_1 and C_2 at our disposal. A basic decision fusion approach to classify one instance would be to naively combine (e.g. concatenate) the data produced by the 3 sources into data ABC, and then classify the instance using C_1 and C_2 on this new combined data. The 2 sets of decisions X_{ABC}^1 and X_{ABC}^2 are then combined with any of the decision fusion methods presented in 2.2 to obtain

the final class label of the instance. This is the strategy we will call *a priori* fusion in this paper. Rather than naively combining data *a priori*, we can also directly use C_1 and C_2 on A, B and C independently, resulting in 6 sets of decisions: X_A^1, X_B^1, X_C^1, X_A^2, X_B^2 and X_C^2, which we will denote $A+B+C$. This is the strategy we will call *a posteriori* fusion in this paper. We could have also only combined two data sources (AB or AC or BC), each resulting in two sets of decisions.

We propose to use all those different approaches simultaneously in the decision fusion process. Rather than feeding a decision fusion method with only 2 sets of decisions on ABC, or 6 sets of decisions on $A+B+C$, we can use $A+B+C+AB+AC+BC+ABC$, resulting in 14 sets of decisions to combine. This approach only exploits available data and does not require to use any additional source.

This strategy is motivated by the idea that some data sources can be more discriminant for certain classes than others. For example, for an action consisting of a hand rotation around the axis of the forearm, the skeleton data will probably not describe the gesture well, because the skeleton is itself the axis of rotation. A gyrometer or accelerometer sensor placed on the hand of the user, on the other hand, will produce data that will feature well the gesture since it will be subject to the rotation of the hand. Conversely, there can be actions that would be easily described with skeleton data, while inertial data would be less useful to classifiers. If we combine *a priori* both of those data sources, the classifiers can use both information to reach better classification performances than when using only one of the sources, but the task of identifying which of the two data is more helpful to classify this specific instance relies only on the classifiers. Using decisions taken on data fused *a priori*, as well as on each data source independently, the task of identifying which data is helpful to classify the instance is shared between the classifiers and the decision fusion method, since it now has access to decisions taken on each data source independently. While the use of decisions taken on each data source independently may worsen the classification rates of the decision fusion method for certain classes that are well featured by all sources, we believe it can significantly improve the rates of classification of the decision fusion method for classes that are only well-captured by some sources, and weakly featured by others.

This process can thus exploit the complementarity of data sources, which will better discriminate classes depending on their nature (e.g. accelerations vs joint trajectories) or their placement (e.g. two accelerometers placed on the hand and on the hip will not equally discriminate a hand rotation).

4 EXPERIMENTAL RESULTS

4.1 Experiments on MHAD

Figure 2: Snapshots from all the actions available in the Berkeley MHAD displayed together with the corresponding point clouds obtained from the Kinect depth data. Actions (from left to right): *jumping*, *jumping jacks*, *bending*, *punching*, *waving two hands*, *waving one hand*, *clapping*, *throwing*, *sit down/stand up*, *sit down*, *stand up*.

The Berkeley MHAD (Multimodal Human Action Database (Ofli et al., 2013), see Figure 2) contains 11 actions performed by 7 male and 5 female subjects in the range 23-30 years of age except for one elderly subject. All the subjects performed 5 repetitions of each action, yielding about 660 action sequences which correspond to about 82 minutes of total recording time.

The specified set of actions consists of the following: (1) actions with movement in both upper and lower extremities (e.g. *Jumping in place*, *Jumping jacks*, *Throwing*, *etc.*), (2) actions with high dynamics in upper extremities (e.g. *Waving hands*, *Clapping hands*, *etc.*), and (3) actions with high dynamics in lower extremities (e.g. *Sit down*, *Stand up*).

4.1.1 Protocols and Method Configurations

We use a *leave-one-subject-out* cross-validation process to test our fusion strategy. Thus, for each cross-validation, all samples from 11 subjects (*i.e.* 605 instances) are used as a training set, and all samples from the remaining subject constitute the test set (*i.e.* 55 instances). A validation set is extracted from 10% of the training set to optimize β of the VK method (it can also be used to optimize certain parameters of other fusion methods, which is not done here for the datasets presented). Thus the training data for the fusion methods only consist of 90% of the actual total training set.

After some preliminary studies on the available 6 accelerometers data and 27 joints skeleton data, we keep 4 main data sources to build our system: the left hand accelerometer (noted A1) and the right hip accelerometer (A4), as in (Chen et al., 2015), as well as the right hand joint trajectory (M20) and the left hand joint trajectory (M27), which are natural joints to track in an action recognition system.

Our fusion method evaluation is based on 3 main action classifiers: DTW, MLP and SNN. The following classifiers and decision fusion methods use the corresponding parameters :

- DTW: low-pass filter of parameter 0.8 on raw data and KNN using K=1;
- MLP: fixed input vectors of size 100 from the concatenation of normalized raw data, 45 hidden neurons with a learning rate of 0.005 and 200 epochs;
- SNN (Berlemont et al., 2015): fixed input vectors of size 100 from the concatenation of normalized raw data, 45 hidden neurons with a learning rate of 0.00005 and 200 epochs;
- VM: $V_c = 1$;
- VBC: $P = 11$ voting classes;
- VW : $V_c = 0.15 + \frac{0.85}{1+e^{-7.68(x-0.68)}}$ and $P = 11$ voting classes;
- VK: β is selected in the set $\{10^{-4}, \ldots, 10^{5}\}$ to maximize classification rates on the validation set;
- SMLP (Hall et al., 2009): 45 hidden neurons, learning rate = 0.2, momentum = 0.1, 100 epochs;
- SSVM (Hall et al., 2009): $C = 256, \gamma = 0.001$;
- PT: $\alpha = 0.5$.

Those parameters were experimentally evaluated (using the same 10% validation set logic) on a purely symbolic gesture database containing accelerometer and gyrometer data only, and were not chosen to be optimal for the MHAD database (or the ChAirGest database presented after).

4.1.2 Inertial and Vision-based Classification

Table 1 presents the average classification rates and standard deviations on isolated data: A1, A4, M20 and M27. It is remarkable to see that for each data source, multiple strategies of decision fusion give better results than the best first-level classifier. For instance, VBC achieves the best results on A1, VM on A4 and M20, and SMLP on M27 data. The best overall classification rate is achieved by VM on M20 with $92.55\% \pm 7.14$, which is better than the best original classifier (SNN) with $91.34\% \pm 6.74$ on M20.

Table 1: Isolated data: average classification rates and standard deviation for standard decision fusion.

	A1	A4	M20	M27
DTW	79.01 ± 10.69	56.54 ± 14.45	88.75 ± 7.55	61.24 ± 8.89
MLP	81.75 ± 9.40	61.10 ± 10.29	89.51 ± 9.27	76.44 ± 11.68
SNN	82.05 ± 9.91	62.76 ± 10.46	91.34 ± 6.74	75.22 ± 10.07
VM	85.39 ± 9.46	$\mathbf{65.49 \pm 10.84}$	$\mathbf{92.55 \pm 7.14}$	75.53 ± 10.56
VBC	$\mathbf{87.81 \pm 9.18}$	61.56 ± 15.04	92.10 ± 7.77	76.44 ± 10.75
VW	85.08 ± 7.60	65.20 ± 10.87	90.42 ± 8.61	76.73 ± 11.39
VK	86.59 ± 9.53	64.90 ± 11.09	92.40 ± 5.61	74.76 ± 9.90
SMLP	84.03 ± 8.09	65.35 ± 12.20	90.58 ± 8.28	$\mathbf{79.78 \pm 10.50}$
SSVM	83.72 ± 8.24	64.58 ± 10.14	90.42 ± 8.11	78.41 ± 11.30
DST	83.72 ± 7.75	62.31 ± 10.51	90.27 ± 8.48	77.04 ± 11.40
PT	83.42 ± 8.15	62.47 ± 10.71	90.42 ± 8.47	76.43 ± 11.67

4.1.3 A priori Data Fusion

Table 2 presents, from concatenated data, the average classification rates and standard deviations of decision fusion methods from 3 classifiers (DTW, MLP, SNN) on A1 and A4 concatenated, on M20 and M27 concatenated, and on A1, A4, M20 and M27 concatenated. This table shows that most decision fusion methods benefit from *a priori* data concatenation. For example, SSVM achieves $87.97\% \pm 7.36$ from A1A4 data which is better than the results it obtained on A1 ($87.81\% \pm 9.18$) or A4 ($65.49\% \pm 10.84$) (see Table 1). The methods that show worse classification rates are actually weakly impacted compared to the results on A1, which shows that decision fusion methods are relatively robust to data that contains non-discriminative parts. The best overall classification rate is achieved by VK on all concatenated data A1A4M20M27 with $96.95\% \pm 3.25$.

Table 2: Concatenated data: average classification rates and standard deviations on *a priori* decision fusion.

	A1A4	M20M27	A1A4M20M27
VM	82.07 ± 8.43	$\mathbf{95.13 \pm 4.87}$	96.64 ± 3.96
VBC	83.73 ± 8.95	94.83 ± 5.14	96.04 ± 3.39
VW	85.39 ± 7.90	93.76 ± 6.12	94.97 ± 5.69
VK	80.86 ± 8.93	94.38 ± 5.61	$\mathbf{96.95 \pm 3.25}$
SMLP	87.52 ± 7.77	93.91 ± 6.42	94.81 ± 5.74
SSVM	$\mathbf{87.97 \pm 7.36}$	93.61 ± 6.58	94.81 ± 5.76
DST	85.85 ± 6.52	93.91 ± 6.45	94.66 ± 6.07
PT	85.40 ± 7.61	94.06 ± 6.37	94.97 ± 5.46

4.1.4 A posteriori Classifier Fusion

Table 3 presents, from separated data, the average classification rates and standard deviations with classifier fusion methods. *A posteriori* decision fusion gives here better average classification rates for each configuration. For instance, when we fuse first-level classifier decisions on separated M20 and M27 data, we obtain $95.90\% \pm 4.32$ for VK, which is higher than $95.13\% \pm 4.87$ presented before (see Table 2) on M20M27 concatenation. The best overall classifica-

tion rate is achieved by VW with $99.39\% \pm 1.11$ on the A1+A4+M20+M27 configuration.

Table 3: Separated data: average classification rates and standard deviations on *a posteriori* decision fusion.

	A1+A4	M20+M27	A1+A4+M20+M27
VM	90.11 ± 6.60	91.35 ± 8.24	98.18 ± 2.01
VBC	89.04 ± 7.62	91.95 ± 7.38	97.71 ± 3.22
VW	92.09 ± 5.63	94.98 ± 5.33	$\mathbf{99.39 \pm 1.11}$s
VK	$\mathbf{92.54 \pm 5.08}$	$\mathbf{95.90 \pm 4.32}$	96.34 ± 3.02
SMLP	91.02 ± 5.79	94.82 ± 5.50	98.32 ± 2.21
SSVM	90.41 ± 5.13	93.45 ± 6.36	98.17 ± 2.32
DST	88.89 ± 8.03	94.08 ± 6.28	96.66 ± 4.35
PT	91.33 ± 5.18	94.68 ± 5.43	98.93 ± 1.67

4.1.5 Mixed Fusion Strategies

Table 4 presents our strategy from mixed data using both *a priori* data and *a posteriori* decision fusion. The best average classification rate is reached at $99.85\% \pm 0.53$ by the VW decision fusion method on separated data (A1+A4+M20+M27), plus partially concatenated data (A1A4+M20M27), plus all concatenated data (A1A4M20M27). These results are significantly better than $99.39\% \pm 1.11$ presented in Table 3.

Table 4: Mixed data: Average classification rates and standard deviations of decision fusion methods from 3 classifiers (DTW, MLP, SNN) on A1, A4, M20 and M27 mixed.

	A1+A4+M20+M27 +A1A4+M20M27	A1+A4+M20+M27 +A1A4+M20M27 +A1A4M20M27
VM	98.48 ± 2.43	98.94 ± 2.12
VBC	98.63 ± 2.93	98.78 ± 2.73
VW	$\mathbf{99.24 \pm 1.20}$	$\mathbf{99.85 \pm 0.53}$
VK	96.80 ± 3.13	97.56 ± 2.85
SMLP	99.24 ± 2.62	99.09 ± 2.63
SSVM	98.78 ± 2.73	99.09 ± 2.63
DST	97.57 ± 4.99	97.87 ± 4.51
PT	98.93 ± 2.74	99.54 ± 1.13

4.1.6 Previous published results

(Chen et al., 2015) propose on this database a fusion approach based on two differing modality sensors (depth and inertial data). Their best results for a *leave-one-subject-out* evaluation is a classification rate of 99.54% fusing Kinect depth stream and A1 and A4 inertial sensors data with a SRC (Sparse Representation Classifier) method. Our strategy is consequently challenging with $99.85\% \pm 0.53$ correct classification at best. We then prove on a second dataset the repeatability of our strategy.

4.2 Experiments on ChAirGest

Figure 3: Two sample images captured by the Kinect RGB stream, where 4 IMUs are fixed on the participant's arm.

The ChAirGest dataset (Ruffieux et al., 2013) contains 6 hours of continuous multi-modal recordings. Data have been acquired from a Kinect camera and 4 Inertial Motion Units (IMUs) attached to the right arm of the subject (see Figure 3). The dataset contains 10 different gestures, started from 3 different resting postures and recorded in two different lighting conditions by 10 different subjects. The 10 gestures considered in the corpus are the following: *Swipe left*, *Swipe right*, *Push to screen*, *Take from screen*, *Palm-up rotation*, *Palm-down rotation*, *Draw a circle I*, *Draw a circle II*, *Wave hello* and *Shake hand*.

4.2.1 Protocols and Method Configurations

As in the previous experiment, we use a *leave-one-subject-out* cross-validation to test our system. Thus, for each cross-validation, all samples from 9 subjects (*i.e.* 450 instances) are used as a training set, and all samples from the remaining subject constitute the test set (*i.e.* 50 instances). Classifiers and fusion methods configurations are identical to the ones used on MHAD, described in section 4.1.1. We use here the 4 accelerometers data (A1, A2, A3 and A4) and 3 joint skeleton data K2, K6, and K10, corresponding respectively to the head, the left hand and the right hand.

4.2.2 Our Results

Table 5: Accelerometer data: Average classification rates and standard deviations.

	A1A2A3A4	A1+A2+A3+A4
VM	76.40 ± 11.23	87.00 ± 6.48
VBC	79.40 ± 9.93	$\mathbf{89.00 \pm 5.35}$
VW	75.80 ± 10.26	87.40 ± 4.90
VK	77.20 ± 10.63	82.20 ± 4.85
SMLP	$\mathbf{81.40 \pm 9.89}$	83.20 ± 6.74
SSVM	73.20 ± 8.95	78.20 ± 7.20
DST	57.40 ± 9.52	79.00 ± 9.20
PT	64.20 ± 7.97	82.80 ± 5.27

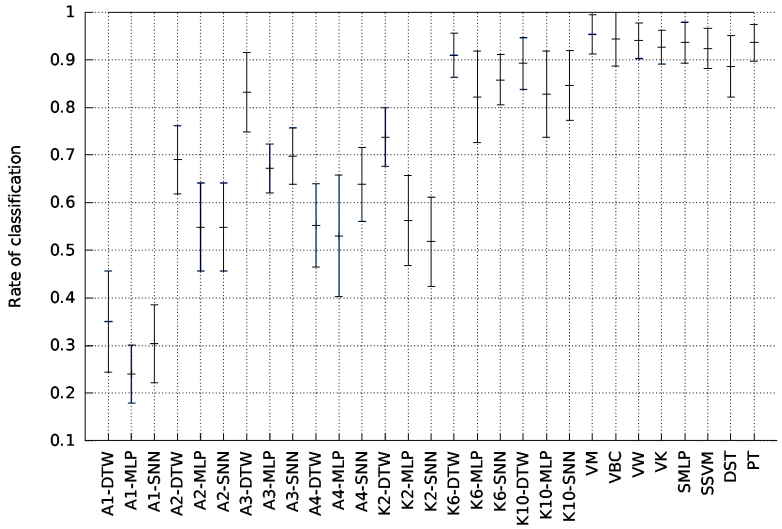

Figure 4: Average classification rates and standard deviations of classifiers on individual data and of decision fusion methods on A1+A2+A3+A4+K2+K6+K10.

Table 5 shows a comparison between *a priori* data and *a posteriori* decision fusion based on 4 accelerometers. The best average classification rate for the second approach is VBC with 89.00% $\pm$ 5.35, as opposed to the first approach where it is SMLP with 81.40% $\pm$ 9.89. We see that each decision fusion method is significantly better on A1+A2+A3+A4 compared to A1A2A3A4, which is a trend that was already existing on the MHAD dataset.

Table 6 proposes similar results on Kinect data with again better classification rates for the second approach (e.g. 94.80% $\pm$ 3.68 for VM) compared to the first one (93.80% $\pm$ 4.05 for VK). However, standard deviations are bigger in the second approach; both approaches are thus quite equivalent for those data.

Figure 4 presents means and standard deviations for each method when applying decision fusion on separated data A1+A2+A3+A4+K2+K6+K10. It is remarkable to see that all decision fusion methods but DST outperform a first-level classifier selection. For instance, the DTW classifier achieves a classification rate of 91.00% $\pm$ 4.64 on K6, while VM achieves 95.40% $\pm$ 4.12 using all separated data sources.

The final results are presented in Table 7. The best overall average classification rate is 96.40% $\pm$ 3.37 for the VM method, with our strategy mixing both *a priori* data and *a posteriori* decision fusion.

4.2.3 Previous published results

(Cao et al., 2015) published results on the ChAirGest dataset. With the exact same *leave-one-subject-out* cross-validation strategy we used in this paper, they attain at best a classification rate of 91.84% $\pm$ 5,76.

Table 6: Kinect data: Average classification rates and standard deviations.

	K2K6K10	K2+K6+K10
VM	93.40 $\pm$ 3.27	**94.80 $\pm$ 3.68**
VBC	93.60 $\pm$ 3.77	93.60 $\pm$ 3.97
VW	93.40 $\pm$ 3.53	93.20 $\pm$ 3.91
VK	**93.80 $\pm$ 4.05**	94.00 $\pm$ 5.33
SMLP	90.40 $\pm$ 3.75	92.00 $\pm$ 3.89
SSVM	89.80 $\pm$ 3.94	91.20 $\pm$ 3.68
DST	90.40 $\pm$ 4.60	90.20 $\pm$ 5.37
PT	90.60 $\pm$ 3.78	92.20 $\pm$ 4.57

Table 7: Mixed data: Average classification rates and standard deviations.

	A1A2A3A4 K2K6K10	A1+A2+A3+A4 +K2+K6+K10	A1+A2+A3+A4 +K2+K6+K10 +A1A2A3A4 +K2K6K10 +A1A2A3A4K2K6K10
VM	89.60 $\pm$ 6.02	**95.40 $\pm$ 4.15**	**96.40 $\pm$ 3.37**
VBC	90.40 $\pm$ 4.88	94.60 $\pm$ 4.62	96.20 $\pm$ 3.58
VW	88.40 $\pm$ 7.35	94.20 $\pm$ 3.82	96.20 $\pm$ 3.46
VK	89.80 $\pm$ 6.14	91.80 $\pm$ 4.62	93.20 $\pm$ 4.02
SMLP	**91.00 $\pm$ 4.45**	93.60 $\pm$ 4.30	95.20 $\pm$ 4.13
SSVM	90.00 $\pm$ 4.11	92.40 $\pm$ 4.20	94.40 $\pm$ 3.75
DST	81.20 $\pm$ 11.04	88.60 $\pm$ 6.47	89.60 $\pm$ 6.31
PT	83.00 $\pm$ 9.39	93.60 $\pm$ 3.86	95.60 $\pm$ 3.62

Our approach performs significantly better for all decision fusion methods, bar DST, with results as high as 96.40% $\pm$ 3.37 for the VM fusion method.

(Yin and Davis, 2013) also published previous results on the ChAirGest dataset. They attain an average final classification rate of 91.16%. Here, our evaluation protocol are quite different (we believe that ours is more challenging), nevertheless our best configuration performs better (96.40% $\pm$ 3.37).

5 CONCLUSIONS AND PERSPECTIVES

In this paper, we tackle the challenge of human action recognition by mixing *a priori* data fusion and *a posteriori* decision fusion. Our strategy applied from 3 main classifiers (DTW, MLP and SNN) on two databases (MHAD (Ofli et al., 2013) and ChAirGest (Ruffieux et al., 2013)) matches or even outperforms state-of-the-art results. Note that the classification rates consistently increased at each step, from standard decision fusion all the way up to our mixed fusion strategy, for all decision fusion methods we studied, which highlights the benefits of our approach.

Our perspectives are to extend our solution to disambiguate some human actions: two gesture classes where frontiers are fuzzy (e.g. *heart* and *clockwise* symbolic gestures); two gesture classes where one source is relevant and the other data source gives no salient information (e.g. limbs rotations, identifiable by inertial systems but not with skeleton based trajectories); or using both strong and weak classifiers, in order to evaluate the impact of extremely low performance classifiers on our approach of decision fusion.

REFERENCES

Akl, A. and Valaee, S. (2010). Accelerometer-based gesture recognition via dynamic-time warping, affinity propagation, & compressive sensing. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Berlemont, S., Lefebvre, G., Duffner, S., and Garcia, C. (2015). Siamese neural network based similarity metric for inertial gesture classification and rejection. *IEEE International Conference on Automatic Face and Gesture Recognition*.

Cao, C., Zhang, Y., and Lu, H. (2015). Multi-modal learning for gesture recognition. In *Multimedia and Expo (ICME), 2015 IEEE International Conference on*, pages 1–6.

Chen, C., Jafari, R., and Kehtarnavaz, N. (2015). Improving human action recognition using fusion of depth camera and inertial sensors. *IEEE Transactions on Human-Machine Systems*, 45(1):51–61.

Cho, S.-J., Choi, E., Bang, W.-C., Yang, J., Sohn, J., Kim, D. Y., Lee, Y.-B., and Kim, S. (2006). Two-stage Recognition of Raw Acceleration Signals for 3-D Gesture-Understanding Cell Phones. In Lorette, G., editor, *Tenth International Workshop on Frontiers in Handwriting Recognition*.

Fauvel, M., Chanussot, J., and Benediktsson, J. A. (2007). *Decision fusion for hyperspectral classification*. John Wiley & Sons, New York, NY, USA.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: An update. In *SIGKDD Explorations*, volume 11.

IBM, Zikopoulos, P., and Eaton, C. (2011). *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. McGraw-Hill Osborne Media.

Kumar, A. and Raj, B. (2015). Unsupervised fusion weight learning in multiple classifier systems. *CoRR arXiv*.

Lefebvre, G., Berlemont, S., Mamalet, F., and Garcia, C. (2015). Inertial gesture recognition with BLSTM-RNN. In *Artificial Neural Networks*, volume 4 of *Springer Series in Bio-/Neuro-informatics*, pages 393–410. Springer International Publishing.

Ofli, F., Chaudhry, R., Kurillo, G., Vidal, R., and Bajcsy, R. (2013). Berkeley MHAD: A comprehensive Multimodal Human Action Database. In *IEEE Workshop on Applications of Computer Vision*, pages 53–60.

Pylvänäinen, T. (2005). Accelerometer Based Gesture Recognition Using Continuous HMMs Pattern Recognition and Image Analysis. volume 3522 of *Lecture Notes in Computer Science*, chapter 77, pages 413–430. Berlin, Heidelberg.

Ruffieux, S., Lalanne, D., and Mugellini, E. (2013). ChAirGest: A Challenge for Multimodal Mid-air Gesture Recognition for Close HCI. In *Proceedings of the 15th ACM International Conference on Multimodal Interaction*, ICMI '13, pages 483–488.

Smets, P. (1990). The combination of evidence in the transferable belief model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(5):447–458.

Vemulapalli, R., Arrate, F., and Chellappa, R. (2014). Human action recognition by representing 3d skeletons as points in a Lie group. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wolpert, D. H. (1992). Stacked generalization. *Neural networks*, 5(2):241–259.

Wu, J., Pan, G., Zhang, D., Qi, G., and Li, S. (2009). Gesture recognition with a 3-d accelerometer. In *Ubiquitous Intelligence and Computing*, volume 5585 of *Lecture Notes in Computer Science*, pages 25–38. Springer Berlin Heidelberg.

Xia, L., Chen, C.-C., and Aggarwal, J. (2012). View invariant human action recognition using histograms of 3d joints. In *IEEE Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 20–27.

Yin, Y. and Davis, R. (2013). Gesture spotting and recognition using salience detection and concatenated hidden markov models. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ICMI '13, pages 489–494.

Zhou, F. and De la Torre Frade, F. (2012). Generalized time warping for multi-modal alignment of human motion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.