A Framework of Multi-Classifier Fusion for Human Action Recognition

Mohammad Ali Bagheri*, Gang Hu*, Qigang Gao*, Sergio Escalera†
* Faculty of Computer Science, Dalhousie University, Halifax, Canada
Email: {bagheri, ghu, qggao}@cs.dal.ca
† Dept. MAIA, University of Barcelona, Gran Via 585, 08007 Barcelona, Spain
Email: sergio@maia.ub.es

***Abstract*—The performance of different action-recognition methods using skeleton joint locations have been recently studied by several computer vision researchers. However, the potential improvement in classification through classifier fusion by ensemble-based methods has remained unattended. In this work, we evaluate the performance of an ensemble of five action learning techniques, each performing the recognition task from a different perspective. The underlying rationale of the fusion approach is that different learners employ varying structures of input descriptors/features to be trained. These varying structures cannot be attached and used by a single learner. In addition, combining the outputs of several learners can reduce the risk of an unfortunate selection of a poorly performing learner. This leads to having a more robust and general-applicable framework. Also, we propose two simple, yet effective, action description techniques. In order to improve the recognition performance, a powerful combination strategy is utilized based on the Dempster-Shafer theory, which can effectively make use of diversity of base learners trained on different sources of information. The recognition results of the individual classifiers are compared with those obtained from fusing the classifiers' output, showing advanced performance of the proposed methodology.**[1]

I. Introduction

The fast and reliable recognition of human actions from captured videos has been a goal of computer vision for decades. Robust action recognition has diverse applications including gaming, sign language interpretation, human-computer interaction (HCI), surveillance, and health care. Understanding gestures/actions from a real-time visual stream is a challenging task for current computer vision algorithms. Over the last decade, spatial-temporal (ST) volume-based holistic approaches and local ST feature representations have been reportedly achieved good performance on some action datasets, but they are still far from being able to express the effective visual information for efficient high-level interpretation. On the other hand, interpreting human actions from tracked body parts is a natural solution that follows the mechanism of human visual perception. The early work conducted by Johansson in 1973 shows that the tracking of joint positions itself encodes significant discriminative information and is sufficient for human beings to recognize different actions [1]. In addition, according to an influential computational model of human visual attention theory [2], visual attention leads to visual salient entities, which provide selective visual information to make human visual perception efficient and effective.

[1]The codes are available at http://web.cs.dal.ca/~bagheri/EnsembleActionRecognition/

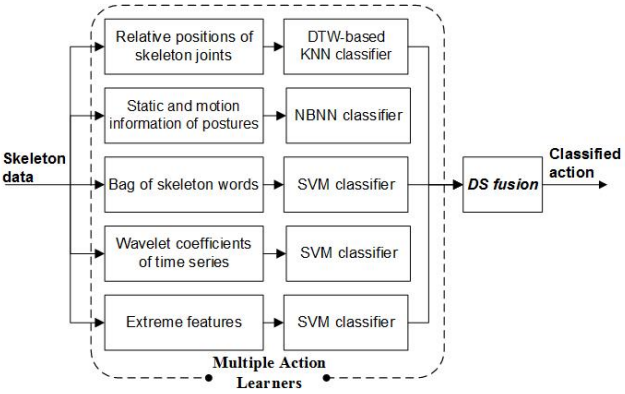

Fig. 1. The framework of the proposed action classification system based on the Dempster-Shafer fusion of multiple classifiers.

Skeleton joints are visual salient points of human body, and their movements in 4D space reflect motion semantics [3]. Development of low-cost depth sensors and evolving skeleton joints detection technique [4] has greatly simplified the task of action recognition, and created a wide range of opportunities for demanding applications. In this paper, we focus on human action recognition by using skeleton joint information extracted from depth sequences.

Action recognition is considered a multi-class classification task where each action type is a separate target class. A classification system involves two main stages: selecting and/or extracting informative features and applying a classification algorithm. In such a system, a desirable feature set can reduce the burden of the classification algorithm, and a powerful classification algorithm can work well even with a low discriminative feature set. In this work, we aim to enhance the efficiency of recognizing human actions by improving both the skeleton-based gesture feature sets and the classification method. In particular, we argue that the discriminative power of skeleton joint information cannot be fully utilized by individual, single recognition techniques. The weakness of single recognition techniques becomes more evident when the complexity of the recognition problem increases, mainly when having many action types and/or similarity of actions. Therefore, we propose the use of an ensemble classification framework in order to improve the efficiency. Fig. 1 shows the framework of our action recognition system, where each combination of a feature set and a classifier is a human

1051-4651/14 $31.00 © 2014 IEEE
DOI 10.1109/ICPR.2014.226

action learner,and the Dempster-Shafer fusion method is used to effectively fuse the outputs of different learners. In this way, the combined efficiency of the ensemble of multiple classification solutions can compensate for a deficiency in one learner. The experimental results show that this strategic combination of these learners can significantly improve the recognition accuracy.

In summary, the contributions of this work are the followings: (1) We apply an ensemble framework to address the action/gesture recognition problem; (2) We efficiently combine individual classifier outputs by means of Dempster-Shafer fusion method, taking benefit from diversity of base classifiers trained on different source of information; (3) We introduce two simple, yet efficient, action description techniques only considering skeleton information.

The rest of the paper is organized as follows: Section 2 reviews the related work on action recognition, and briefly introduces multiple classifiers systems. Section 3 presents the framework of our multi-classifier fusion for action recognition. Section 4 evaluates the proposed system. Finally, Section 5 is the conclusion.

II. Related Work

A. Action Recognition

Various representational methodologies have been proposed to recognize human actions/gestures. Based on extracted salient points or regions [5], [6] from ST volume, several local ST descriptor methods, such as HOG/HOF [7] and extended SURF [8] have been widely used for human action recognition from RGB data. Inspired from the text mining area, the intermediate level feature descriptor for RGB videos, Bag-of-Word (BoW)[9], [10], has been popular due to its semantic representation and robustness to noise. Recently, BoW-based methods have been extended to depth data. In [11], Bag-of-Visual-and-Depth-Words defined containing a vocabulary from RGB and depth sequences. This novel representation was also used to perform multi-modal action recognition.With the development of low-cost depth sensors, Shotton et al. [4] proposed a Random Forest-based classification method to find body joints from depth images in an efficient way. This approach has been recently enhanced to provide accurate 3D estimations of skeleton joint locations.

Real time skeleton data facilitates the human activity analysis research. In [12], visual features for activity recognition are computed based on the spatial and temporal differences among detected joints. This feature set contains information about static posture, motion, and offset. Then, Naive Bayes Nearest Neighbor method was applied for the classification task. Alternatively, a histogram of 3-D joint locations (HOJ3-D) for body posture representation is proposed in [13]. In this representation, the 3D space is partitioned into bins using a spherical coordinate system, and the HOJ3-D histogram is constructed by casting joints into certain bins. After applying linear discriminant analysis (LDA) for dimensionality reduction, HOJ3-D vectors are clustered into k posture visual words. The temporal behaviour of these visual words is coded by discrete HMMs. Wang et al. [14] described skeleton joints by pairwise and local occupancy patterns, and used Fourier Temporal Pyramid to model the temporal patterns of the joint feature vectors. Their Actionlet Ensemble (AE) model can handle errors of the skeleton tracking and better characterize the intra-class variations. Despite active research for action /gesture recognition, none of the previous skeleton-based approaches considers a multiple classifier system philosophy.

B. Multiple Classifier Systems

The efficiency of pattern classification by a single classifier has been recently challenged by multiple classifier systems [15], [16], [17]. A multiple classifier system is a classification system made up of an ensemble of individual classifiers whose outputs are combined in some way to obtain a final classification decision. In an ensemble classification system, it is hoped that each base classifier will focus on different aspects of the data and will err under different situations [16]. However, the ensemble approach depends on the assumption that single classifiers' errors are uncorrelated, which is known as classifier *diversity* in the background literature [18]. The intuition is that if each classifier makes different errors, then the total errors can be reduced by an appropriate combination of these classifiers.

Once a set of classifiers is generated, the next step is to construct a combination function to merge their outputs, which is also called decision optimization. The most straightforward strategy is the simple majority voting, in which each classifier votes on the class it predicts, and the class receiving the largest number of votes is the ensemble decision. Other strategies for combination function include weighted majority voting, sum, product, maximum and minimum, fuzzy integral, decision templates, and the Dempster-Shafer (DS) based combiner [19],[15]. Inspired by the Dempster-Shafer (DS) theory of evidence [20], a combination method is proposed in [21], which is commonly known as the Dempster-Shafer fusion method. By interpreting the output of a classifier as a measure of evidence provided by the source that generated the training data, the DS method fuses an ensemble of classifiers [22].

In this work, after defining a set of skeleton-based feature spaces, we train different action/gesture recognition models whose outputs are fused based on the DS fusion algorithm. As a result, we show that we can merge predictions made from different learners, trained in different feature spaces, with different dimensionality in both feature space and action/gesture sample length. Following the multiple classifiers philosophy, we show that the proposed ensemble approach outperforms standard non-ensemble strategies for action recognition.

III. Framework of Multi-Classifier Fusion For Action Recognition

Here we present a framework of multi-classifier fusion for human action recognition. First, individual skeleton data-based human action classifiers, so-called action learners are introduced. Then, we describe the techniques applied in this system, in particular the fusion mechanism employed.

A. Human Action Learners

In this work, we have employed five different action recognition techniques using only positions of skeleton joints, described in the following subsections. Among them, the first three methods are the existing ones [12],[23],[24] in the

related literature and the last two methods are our proposed techniques.

1) EigenJoints + Naive-Bayes-Nearest-Neighbor
In [12], visual features for activity recognition are computed based on the spatial and temporal differences between skeleton joints, named *EigenJoints* features. This feature set contains information about static posture, motion, and offset. In their method, a feature descriptor is generated for each frame. Therefore, each action has a different number of feature descriptors, depending on the number of frames. For classification, they employed the Naive Bayes Nearest Neighbor (NBNN) method, which is a very efficient method recently introduced for image classification [25].

2) Dynamic Time Warping + KNN
Dynamic Time Warping (DTW) is a well-known algorithm which aims to compare and align two temporal sequences, taking into account that sequences may vary in length (time) [23]. DTW employs the dynamic programming technique to find the minimal distance between two time series, where sequences are warped by stretching or shrinking the time dimension. Although it was originally developed for speech recognition [26], it has also been employed in many other areas like handwriting recognition, econometrics, and action recognition.

In this work, depending on the problem, the relative distance of suitable joint points is obtained at each frame. Then, given two actions represented by two multi dimensional time series, DTW calculates the distance between two actions. To classify an unlabeled test action (sample), its distance to all training samples is calculated. Consequently, the nearest neighbor algorithm should be employed for classification. Given a test action, we calculate its distance to all training actions using DTW, and the target of the closest sample is predicted as the target class.

3) Bag of skeleton words + SVM
We also apply the BoW approach to 3D joint data [11], [24], such that each visual word is constructed using a set of spatio-temporal descriptors of skeleton joint positions. In fact, each visual word, conceptually, is the cluster centre and represents a unique posture. We further choose words, i.e. clusters, with high discrimination capability. To this end, we compute the entropy of each word using the distribution of class samples in the corresponding cluster, and select the words with top half entropy. As an example, an initial word might represent the neutral posture. Since each action usually includes frames showing the neutral posture, this word will be removed from the dictionary. Then, each action is represented using the frequency of each word in the codebook, obtaining a histogram of words for each action. These histograms can be used as the input features for a particular classifier, e.g. SVM classifier is used here.

For the next two sets of our proposed action descriptors, i.e. wavelet coefficients of time series, and extreme features, all standard classifiers can be used in the classification stage. Here, we chose Support Vector Machine (SVM) with the Gaussian Kernel as the classification algorithm, since SVM is a state-of-the-art classifier and is commonly used for image and video recognition.

4) Wavelet coefficients + SVM
Here, we propose a new technique for action description by using the multilevel wavelet decomposition technique based on the Mallat algorithm [27]. The Mallat algorithm is a classical scheme, known as two-channel subband coding in the signal processing community. When a signal passes through the filters, the low frequency components ,often called approximation in wavelet theory, and high frequency components (details) are emerged.The low frequencies are usually the most important part of a signal and represent its identity. The decomposition process can be repeated on the approximation components, so that one signal is broken down into many lower resolution components. This procedure is known as multilevel discrete wavelet transform (DWT) or simply multilevel decomposition.

In the context of action recognition using skeleton joint information, the relative position of each skeleton joint during the time is considered as a signal (time series). Then, we apply the multilevel wavelet decomposition method to extract low level wavelet coefficients.

In our implementation, three series are generated for each skeleton joint, according to the three dimensions of the real world position of the joint (X, Y, Z). The final feature vector for each action is generated by concatenating the wavelet coefficients of all time series.

5) Extreme features + SVM
We also propose a simple, but very effective, action description technique. This method is based on the idea that for many short actions, like those in the benchmark datasets, only a very few salient postures can be a unique representative of the action. These postures are unique in the sense that the relative position, i.e. the distance, between a set of skeleton joints will reach its extreme value.

In this method, given an action, we compute the pair distance between each appropriate joint point at each frame. The feature vector is then generated by taking the maximum and minimum value of each pair distance of all frames:

$$ExtremeFeatures = \bigcup \{min_t(PD^t_{ijk}), max_t(PD^t_{ijk})\} \quad (1)$$

where PD^t_{ijk} is the pairwise distance of ith and jth joint points during the time in the kth dimension (i.e. x, y, z). Depending on the problem, we may use all available points, e.g. the 20 joint points from Kinect APIs, or a subset of appropriate points. For each joint pair, six features will be generated, which are the maximum and minimum distances of two joints in X, Y, and Z dimensions.

B. Dempster-Shafer fusion method

Let $x \in R^n$ be a feature vector and $\Omega = \{\omega_1, \omega_2, \ldots, \omega_c\}$ be the set of class labels. Each classifier h_i in the ensemble $H = \{h_1, h_2, \ldots, h_L\}$ outputs c degrees of support. Without loss of generality, we can assume that all c degrees are in the interval $[0, 1]$. The support that classifier h_i, gives to the hypothesis that **x** comes from class ω_j is denoted by $d_{i,j}(x)$. Clearly, the larger the support, the more likely the class label ω_j. The L classifier outputs for a particular instance **x** can be organized in a decision profile, $DP(x)$, as the following

matrix [15]:

$$DP(x) = \begin{pmatrix} d_{1,1}(x) & \cdots & d_{1,j}(x) & \cdots & d_{1,c}(x) \\ \vdots & & \vdots & & \vdots \\ d_{i,1}(x) & \cdots & d_{i,j}(x) & \cdots & d_{i,c}(x) \\ \vdots & & \vdots & & \vdots \\ d_{L,1}(x) & \cdots & d_{L,j}(x) & \cdots & d_{L,c}(x) \end{pmatrix}$$

The Dempster-Shafer fusion method uses decision profile to find the overall support for each class and subsequently labels the instance **x** in the class with the largest support. In order to obtain the ensemble decision based on DS fusion method, first, the c decision templates, $DT_1, \ldots, DT_c$, are built from the training data. Roughly speaking, decision templates are the most typical decision profile for each class ω_j. For each test sample, **x**, the DS method compare the decision profile, $DP(x)$, with decision templates. The closest match will label **x**. In order to predict the target class of each test sample, the following steps are performed [15][21]:

1. Build decision templates: For $j = 1, \ldots, c$, calculate the means of the decision profiles for all training samples belonging to ω_j. Call the mean a decision template of class ω_j, DT_j.

$$DT_j = \frac{1}{N_j} \sum_{z_k \in \omega_j} DP(z_k) \tag{2}$$

where N_j in the number of training samples belong to ω_j.

2. Calculate the proximity: Let DT_j^i denote the ith row of the decision template DT_j, and D_i the output of the ith classifier, that is, the ith row of the decision profile $DP(x)$. Instead of similarity, we now calculate proximity Φ, between DT_j^i and the output of classifier D_i for the test sample x:

$$\Phi_{j,i}(x) = \frac{(1 + \|DT_j^i - D_i(x)\|)^{-1}}{\sum_{k=1}^{c}(1 + \|DT_j^i - D_i(x)\|)^{-1}} \tag{3}$$

where $\|.\|$ is a matrix norm.

3. Compute belief degrees: Using Eq. (2), calculate for each class $j = 1, \ldots, c$ and for each classifier $i = 1, \ldots, L$, the following belief degrees, or evidence, that the ith classifier is correctly identifying sample **x** into class ω_j:

$$b_j(D_i(x)) \frac{\Phi_{j,i}(x) \prod_{k \neq j}(1 - \Phi_{k,i}(x))}{1 - \Phi_{j,i}(x)[1 - \prod_{k \neq j}(1 - \Phi_{k,i}(x))]} \tag{4}$$

4. Final decision based on class support: Once the belief degrees are achieved for each source (classifier), they can be combined by Dempster's rule of combination, which simply states that the evidences (belief degree) from each source should be multiplied to obtain the final support for each class:

$$\mu_j(x) = K \prod_{i=1} b_j(D_i(x)), \ j = 1, \ldots, c$$

where K is a normalizing constant ensuring that the total support for ω_j from all classifiers is 1. The DS combiner gives a preference to class with largest $\mu_j(x)$.

IV. Experiments

A. Datasets

We evaluated our framework on two publicly available datasets: the Multi-modal Gesture Recognition Challenge 2013 (Chalearn) and MSR Action3D.

Chalearn dataset: This dataset is a newly released large video database of 13,858 gestures from a lexicon of 20 Italian gesture categories recorded with a Kinect camera, including audio, skeletal model, user mask, RGB and depth images [28]. It contains image sequences capturing 27 subjects performing natural communicative gestures and speaking in fluent Italian, and is divided into development, validation and test parts. We conducted our experiments on the depth images of development and validation samples which contains 11,116 gestures across over 680 depth sequences. Each sequence lasts between 1 and 2 minutes and contains between 8 and 20 gesture samples, around 1,800 frames. Some examples of RGB images are shown in Fig. 2.

MSR-Action3D dataset: This dataset [24] is a well-known benchmark dataset for 3D action recognition. This dataset contains 20 actions, including *high arm wave, horizontal arm wave, hammer, hand catch, forward punch, high throw, draw x, draw tick, draw circle, hand clap, two hand wave, side-boxing, bend, forward kick, side kick, jogging, tennis swing, tennis serve, golf swing, pick up & throw*. Each action was performed 2 or 3 times by each subject. Skeleton joint data of each frame is available having a variety of motions related to arms, legs, torso, and their combinations. In total, there are 567 depth map sequences with a resolution of 320 $\times$ 240.

B. Classification Results

For Chalearn dataset, the classification performance is obtained by means of stratified 5-fold cross-validation. For MSR Action3D dataset, most studies follow the experimental setting of Li et al. [24], such that they first divide the 20 actions into three subsets, each having 8 actions. For each subset, they perform three tests. In test one and two, 1/3 and 2/3 of the samples were used as training samples and the rest as testing samples. In the third test, half of the subjects are used as training and the rest subjects as testing. The experimental results on the first two tests are generally very promising, mainly more than 90% accuracy. On the third test, however, the recognition performance dramatically decreases. It shows that many of these methods do not have good generalization ability when a different subject is performing the action, even in the same environmental settings. In order to have more reliable results, we followed the same experimental setup of [14], [29]. In this setting, actors 1,3,5,7, and 9 are used for training and the rest for testing.

As mentioned before, we chose SVM with the Gaussian Kernel as the base classifier for the last three sets of action descriptors. Also, dynamic time warping is a distance-based method and therefore we employed the nearest neighbor algorithm for classification. In addition, in the method proposed in [12], each action has different number of feature descriptors. Therefore, standard classifiers, like SVM or neural networks, cannot be used as the classifier. Similar to their paper, we implemented and used the NBNN classifier.

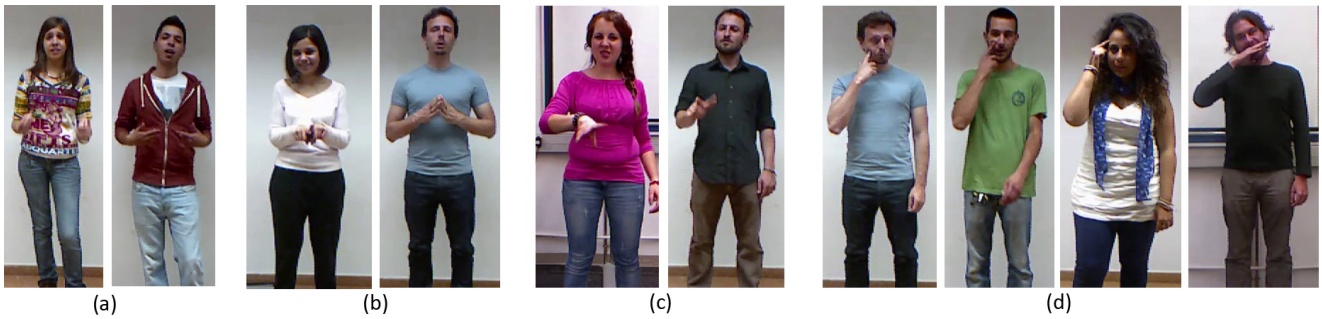

Fig. 2. Some example gestures in the Chaleran dataset are very easy to be confused, even from human visual perception.
(a) *Che vuoi* vs. *Che due palle*. For the *Che vuoi* gesture, both hands are in front of the chest area; where for *Che due palle* gesture they are near the waist region. (b) *Vanno d'accordo* vs. *Cos hai combinato*: both hand positions are very close and with the same motion directions; (c) both gestures, *Si sono messid'accordo* and *non ce ne piu*, require hand rotations; (d) four gestures, *Furbo, seipazzo, buonissimo*, and *cosatifarei* are required with the finger pointing to the head area, which cannot be easily determined, even with human eyes.

TABLE I. CLASSIFICATION ACCURACY OF INDIVIDUAL ACTION LEARNERS ON THE CHALEARN GESTURE DATASET.

	EigenJoints +NBNN	**DTW +KNN**	**BoVW +SVM**	**Wavelet Coeff +SVM**	**Extreme Features +SVM**	**DS-Fusion**
5 classes	63.50	97.40	95.40	91.60	96.80	**99.12**
10 classes	58.70	89.80	88.00	86.60	88.40	**93.78**
15 classes	56.17	86.82	85.27	79.64	87.00	**90.43**
20 classes	54.30	77.85	73.05	70.40	73.65	**82.60**
Average	58.17	87.97	85.43	82.06	86.46	**91.48**

TABLE II. CLASSIFICATION ACCURACY OF INDIVIDUAL ACTION LEARNER AND THE FUSED CLASSIFIER ON THE MSR ACTION3D DATASET.

	EigenJoints +NBNN	**DTW +KNN**	**BoVW +SVM**	**Wavelet Coeff +SVM**	**Extreme Features +SVM**	**DS-Fusion**
5 classes	72.97	95.95	83.78	81.62	96.95	**99.15**
10 classes	47.62	85.14	82.43	85.14	87.84	**89.92**
15 classes	44.14	82.60	77.30	63.19	78.46	**83.11**
20 classes	47.81	78.76	64.65	58.92	72.39	**84.85**
Average	53.14	84.61	77.04	72.22	83.91	**89.26**

The summaries of the results are reported in Table I and Table II for Chalearn and MSR Action3D datasets. In addition, the confusion matrix of the ensemble classification system for both datasets are demonstrated in Figure 3. It is important to note the outperformance of the fused results in comparison with the individual classifier. The result are quite promising, considering the facts that the skeleton tracker sometimes fails and the tracked joint positions are quite noisy.In these tables, the effectiveness of the proposed action description technique based on the extreme pair distance of joint points is notable. For both considered datasets, our fast method achieved very high accuracy in many cases.

We then compare our ensemble classification results on MSR Action3D dataset with state-of-the-art methods. Table III shows the accuracy of our method and the rival methods on this dataset based on the cross-subject test setting [24]. Some of the methods in this table, like HOG-3D [32] and HON4D [29], use depth data in addition to skeleton joint information. However, processing sequences of depth images is much more computationally intensive. Even though the accuracy of the proposed framework is slightly less than HON4D [29], the advantage of our method is its fast implementation, which makes it feasible for real-time applications.

TABLE III. COMPARING CLASSIFICATION ACCURACY OF OUR ENSEMBLE FRAMEWORK WITH THE STATE-OF-THE-ART METHODS ON THE MSR-ACTION3D DATASET.

Method	**Accuracy**
Recurrent Neural Network [30]	42.5
Hidden Markov Model [31]	54
Action Graph on Bag of 3D Points [24]	74.7
HOG 3D [32]	81.43
HON4D [29]	85.85
Dollar + BOW [8]	72.40
STIP [5] + BOW	69.57
Vieira et al. [33]	78.20
Our method	**84.85**

V. CONCLUSION

This paper presents an ensemble classification framework to address the multiple action/gesture recognition problem. We designed a set of classifiers, each one trained over different feature sets. We focused on feature spaces defined by a 3D skeletal model of the human body and proposed two simple, yet effective, feature representations for this problem. The overall performance of the ensemble of classifiers is improved by fusing the classifiers using the Dempster-Shafer combination theory. We compared the classification results of the individual classifiers with those obtained from fusing the classifiers by the Dempster-Shafer combination method on two public datasets, showing significant performance improvements of the proposed methodology. Considering the fact that we have only employed the position of skeleton joints, our accuracy results on MSR-Action3D dataset have reached the highest level comparing with other state-of-the-arts methods. In conclusion, we found that using ensemble methods for human actions and gestures classification is an effective approach.

REFERENCES

[1] G. Johansson, "Visual perception of biological motion and a model for its analysis," *Perception & psychophysics*, vol. 14, no. 2, pp. 201–211, 1973.

[2] A. Treisman and H. Schmidt, "Illusory conjunctions in the perception of objects," *Cognitive psychology*, vol. 14, no. 1, pp. 107–141, 1982.

[3] G. Hu and Q. Gao, "A 3d gesture recognition framework based on hierarchical visual attention and perceptual organization models," in *Pattern Recognition (ICPR), 2012 21st International Conference on*. IEEE, 2012, pp. 1411–1414.

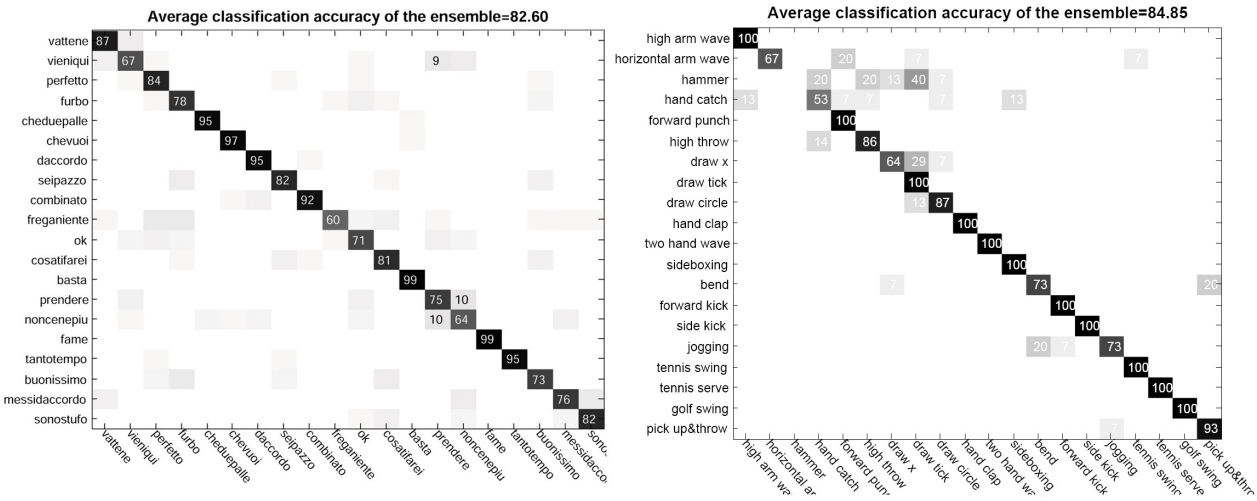

Fig. 3. Confusion matrices of the ensemble classification system on the Chalearn (left) and MSR-Action3D datasets (right).

[4] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[5] I. Laptev, "On space-time interest points," *IJCV*, vol. 64, no. 2-3, pp. 107–123, 2005.

[6] G. Willems, T. Tuytelaars, and L. Van Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," in *ECCV*. Springer, 2008, pp. 650–663.

[7] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.

[8] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *IEEE Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*. IEEE, 2005, pp. 65–72.

[9] J. Liu, B. Kuipers, and S. Savarese, "Recognizing human actions by attributes," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2011, pp. 3337–3344.

[10] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2011, pp. 3169–3176.

[11] A. Hernndez-Vela, M. A. Bautista, X. Perez-Sala, V. Ponce, X. Bar, O. Pujol, C. Angulo, and S. Escalera, "Bovdw: Bag-of-visual-and-depth-words for gesture recognition," in *ICPR*. IEEE, 2012, pp. 449–452.

[12] X. Yang and Y. Tian, "Eigenjoints-based action recognition using naive-bayes-nearest-neighbor," in *CVPR Workshops (CVPRW)*. IEEE, 2012, pp. 14–19.

[13] L. Xia, C.-C. Chen, and J. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *CVPR Workshops (CVPRW)*. IEEE, 2012, pp. 20–27.

[14] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 1290–1297.

[15] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. New York, NY: Wiley, 2004.

[16] R. Polikar, "Ensemble based systems in decision making," *IEEE Circuits and Systems Magazine*, vol. 6, no. 3, pp. 21–45, 2006.

[17] M. A. Bagheri, Q. Gao, and S. Escalera, "A framework towards the unification of ensemble classification methods," in *12th International Conference on Machine Learning and Applications (ICMLA)*, vol. 2, 2013, pp. 351–355.

[18] T. Windeatt, "Accuracy/diversity and ensemble mlp classifier design," *IEEE Transactions on Neural Networks*, vol. 17, no. 5, pp. 1194–1211, 2006.

[19] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Inteligence*, vol. 20, no. 3, p. 226239, 1998.

[20] A. Dempster, "Upper and lower probabilities induced by multivalued mappings," *Annals of Mathematical Statistics*, vol. 38, no. 2, pp. 325–339, 1967.

[21] G. Rogova, "Combining the results of several neural network classifiers," *Neural Networks*, vol. 7, pp. 777–781, 1994.

[22] M. Bagheri, Q. Gao, and S. Escalera, "Logo recognition based on the dempster-shafer fusion of multiple classifiers," in *26th Canadian conf. on Artificial Intelligence*, Regina, Canada, 2013.

[23] M. Reyes, G. Dominguez, and S. Escalera, "Feature weighting in dynamic timewarping for gesture recognition in depth data," in *CVPR Workshops (CVPRW)*. IEEE, 2011, pp. 1182–1188.

[24] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3d points," in *CVPR Workshop (CVPRW)*. IEEE, 2010, pp. 9–14.

[25] O. Boiman, E. Shechtman, and M. Irani, "In defense of nearest-neighbor based image classification," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.

[26] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 26, no. 1, pp. 43–49, 1978.

[27] S. Mallat, *A Wavelet Tour of Signal Processing*, 2nd ed. San Diego, CA: Academic, 1999.

[28] S. Escalera, J. Gonzlez, X. Bar, M. Reyes, O. Lopes, I. Guyon, V. Athistos, and H. Escalante, "Multi-modal gesture recognition challenge 2013: Dataset and results," in *ICMI*, 2013.

[29] O. Oreifej, Z. Liu, and W. Redmond, "Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[30] J. Martens and I. Sutskever, "Learning recurrent neural networks with hessian-free optimization," in *ICML*, 2011, pp. 1033–1040.

[31] F. Lv and R. Nevatia, "Recognition and segmentation of 3-d human action using hmm and multi-class adaboost," in *ECCV*. Springer, 2006, pp. 359–372.

[32] A. Klaser and M. Marszalek, "A spatio-temporal descriptor based on 3d-gradients," in *British Machine Vision Conference*, 2008.

[33] A. W. Vieira, E. R. Nascimento, G. L. Oliveira, Z. Liu, and M. F. Campos, "Stop: Space-time occupancy patterns for 3d action recognition from depth map sequences," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*. Springer, 2012, pp. 252–259.