Keep it Accurate and Diverse: Enhancing Action Recognition Performance by Ensemble Learning

Mohammad Ali Bagheri, Qigang Gao
Faculty of Computer Science, Dalhousie University, Halifax, Canada
bagheri@cs.dal.ca

Sergio Escalera, Albert Clapes
Computer Vision Center, UAB
Edificio O, Campus UAB, 08193, Bellaterra (Cerdanyola), Barcelona, Spain
Dept. Applied Methematics, University of Barcelona
Gran Via de les Corts Catalanes, 585, 08007, Barcelona
sergio@maia.ub.es, aclapes@cvc.uab.cat

Kamal Nasrollahi, Michael B. Holte, and Thomas B. Moeslund
Visual Analysis of People (VAP) Laboratory
Rendsburggade 14, 9000 Aalborg, Denmark
{kn, mbh, tbm}@create.aau.dk

Abstract

The performance of different action recognition techniques has recently been 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 action learning techniques, each performing the recognition task from a different perspective. The underlying idea is that instead of aiming a very sophisticated and powerful representation/learning technique, we can learn action categories using a set of relatively simple and diverse classifiers, each trained with different feature set. In addition, combining the outputs of several learners can reduce the risk of an unfortunate selection of a learner on an unseen action recognition scenario. This leads to having a more robust and general-applicable framework. 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 enhanced performance of the proposed methodology.

1. Introduction

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 [1]. 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.

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 the classification module. In particular, we argue that the discriminative power of encoded information can-

not 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, where each combination of a feature set and a classifier is a human 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.

It is worth mentioning that there are two general approaches to employ the power of different representation/description techniques. First approach is to concatenate the obtained feature sets (early fusion), and then to fed this higher dimensional feature set to a single classifier. The second approach, which has not been fully employed, is to train different individual classifiers; each trained on a feature set, and then efficiently combined in a late fusion fashion.

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.

2. Related work

2.1. Action recognition

Various representational methodologies have been proposed to recognize human actions/gestures. Based on extracted salient points or regions [14, 32] from ST volume, several local ST descriptor methods, such as HOG/HOF [15] and extended SURF [7] 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-Visual-Word (BoVW)[16, 29], has been developed due to its semantic representation and robustness to noise. Recently, BoVW-based methods have been extended to depth data. In [8], Bag-of-Visual-and-Depth-Words defined to contain a vocabulary from RGB and depth sequences. This novel representation was also used to perform multi-modal action recognition.

Low-level local features are popular for representing video information. State-of-the-art performance for large scale action recognition has been achieved when combined with a Bag-of-Words (BoVW) or Fisher vector feature representation, and linear or non-linear Support Vector Machine (SVM) classification [30]. A recent evaluation by Wang et al. [31] has shown how dense feature sampling improves performance over local feature description of the neighborhood of sparse interest points for action recognition.

Some of the most popular low-level features descriptors are the Histograms of Oriented Gradients (HOG), the Histograms of Optical Flow (HOF), the Histograms of Oriented 3D spatio-temporal Gradients (HOG3D), and the Motion boundary Histograms (MBH) descriptor, yielding remarkable results on a variety of datasets for action recognition in comparison with other state-of-the-art descriptors. HOG captures the static appearance (gradient) information and HOF captures the local motion (flow) information [15]. The HOG3D descriptor [12] is a spatio-temporal volume representation of gradients and generalizes the HOG concepts to 3D. The MBH descriptor were proposed by Dalal et al. [5] for human detection by computing derivatives separately for the horizontal and vertical components of the optical flow. The descriptor encodes the relative motion between pixels, locally constant camera motion is removed and information about changes in the motion boundaries is kept, resulting in more robustness to camera motion than optical flow.

Wang and Schmid proposed an approach based on the recent improvement of low-level dense trajectory features [30]. They extract local HOG, HOF and MBH feature descriptors from dense trajectories, and apply Fisher vectors to integrate them into a compact representation for each video. Finally, a linear SVM with one-against-rest is employed for the multi-class action classification problem. using a histogram intersection kernel. Hence, these methods are based on combination of multiple low-level features, where the feature extraction involves cuboid computation in a 3D spatio-temporal space.

2.2. Multiple classifier systems

The efficiency of pattern classification by a single classifier has been recently challenged by multiple classifier systems [13, 24]. 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 [24]. However, the ensemble approach depends on the assumption that single classifiers' errors are uncorrelated, which is known as classifier *diversity* in the background literature [33]. 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 [11],[13]. Inspired by the Dempster-Shafer (DS) theory of evidence [6], a combination method is proposed in [25], 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.

In this work, after extracting a set of visual feature sets, we train different action learning 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 sample length. Following the multiple classifiers philosophy, we show that the proposed ensemble approach outperforms standard non-ensemble strategies for action recognition.

2.3. 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 [13]:

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

The Dempster-Shafer fusion method uses a 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 [13][25]:

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.

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

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:

$$\Phi_{j,i}(x) = \frac{(1 + \|DT_j^i - D_i(x)\|)^{-1}}{\sum_{k=1}^{c}(1 + \|DT_j^i - D_i(x)\|)^{-1}} \tag{2}$$

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:

$$b_j(D_i(x)) \frac{\Phi_{j,i}(x) \prod_{k \neq j}(1 - \Phi_{k,i}(x))}{1 - \Phi_{j,i}(x)[1 - \prod_{k \neq j}(1 - \Phi_{k,i}(x))]} \tag{3}$$

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 $\mu_j(x)$.

3. Three approaches to action recognition

In this paper, we have utilized three different approaches to recognize action categories. The first approach, is considered as a baseline for comparing with the second and third ensemble-based approaches.

Approach 1: The straightforward approach to utilize the five extracted feature sets is to combine them and generate

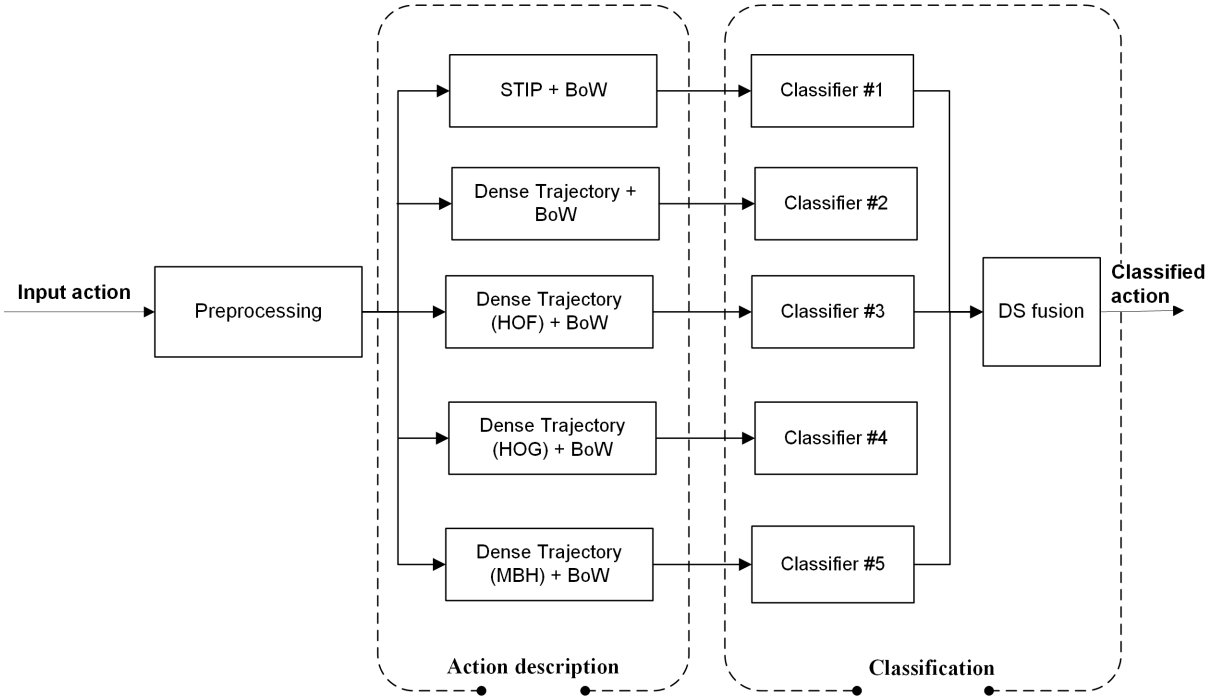

Figure 1. The framework of the proposed action classification system based on the Dempster-Shafer fusion of multiple classifiers.

a higher dimensional feature set. This feature set is fed into a single classifier to recognize action classes.

Approach 2: In this approach, shown in Figure 1, different action representation techniques are first employed to describe action samples. Then, each feature set is fed into its corresponding classifier (learner); therefore an ensemble of classifiers is generated. The outputs of these classifiers are finally fused using an efficient combination method.

Approach 3: The third approach follows the underlying idea of the Random Subspace Method (RSM)[9], in which each classifier in the ensemble is trained using a random subset of features. In this work, however, we first concatenated the five above mentioned feature sets. Then, a number of feature subsets are randomly chosen from all features; and then each feature set are used to train an individual classifier. The outputs of ensemble of classifiers are then combined.

For all approaches, first, we have applied five different action description techniques: the well-known space-time interest points (STIPs) method and four description techniques based on the dense trajectory works of Wang et al. [29, 30], including dense Trajectory, dense trajectories of HOF descriptors, dense trajectories of HOG descriptors, and dense trajectories of MBH descriptors. It is worth mentioning that extracted features are encoded using conventional Bag-of-Words technique, resulting in histograms of 4000 bins. Therefore, five individual feature sets are generated, each having 4000 features for each sample. In this work, concatenating five action description sets will build a 20,000 dimensional feature set.

4. Experiments

4.1. Dataset

We evaluated the proposed approach on the UCF101 dataset, an extension of the UCF-50 dataset. UCF101 [27] is a large action recognition database of real action videos, collected from YouTube. The dataset consists of 13,320 videos belonging to 101 categories that are separated into 5 broad groups: Human-Object interaction (applying eye makeup, brushing teeth, hammering, etc.), Body-Motion (Baby crawling, push ups, blowing candles, etc.), Human-Human interaction (Head massage, salsa spin, haircut, etc.), Playing Instruments (flute, guitar, piano, etc.), and Sports, as shown in Figure 2. Having 13,320 videos from 101 categories gives the largest diversity in terms of actions and with the presence of large variations in camera motion, object appearance and pose, object scale, viewpoint, cluttered background, illumination conditions, etc., making it one of the most challenging datasets to date for action recognition.

4.2. Experimental settings

To divide 13,320 instances into train and test sets, we followed the division procedure proposed by the authors of UCF101 [27], such that instances are divided into three training/testing splits, where videos for each of the 25 groups remain separated. For classification, we chose SVM

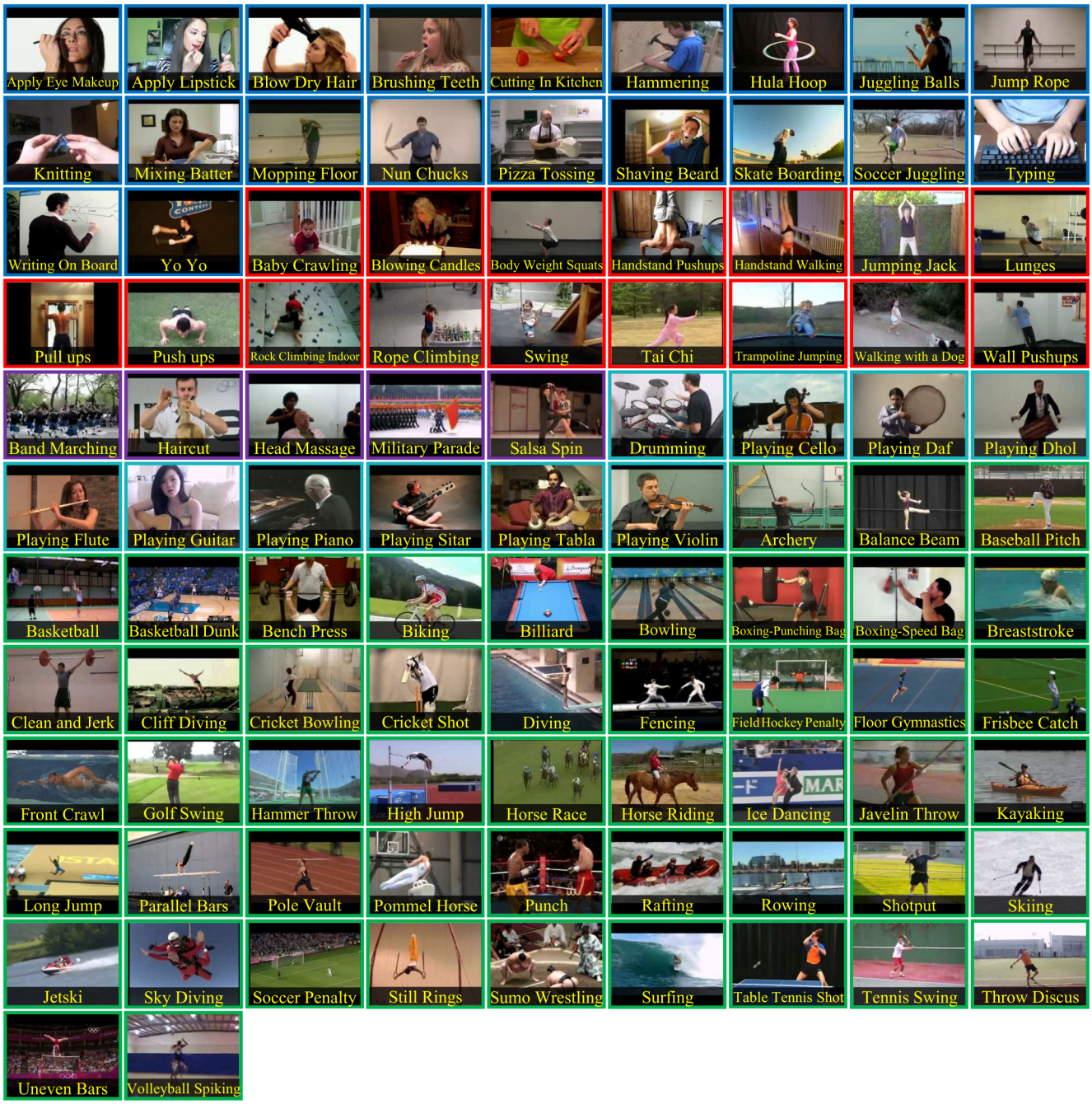

Figure 2. The 101 actions categories included in the UCF101 dataset shown with a single frame. The color of the frame borders corresponds to which of the five action types the action samples belong: (1) Human-Object Interaction, (2) Body-Motion Only, (3) Human-Human Interaction, (4) Playing Musical Instruments, (5) Sports [27].

with the histogram intersection kernel [17] as the base classifiers.

4.3. Classification results

The summaries of the obtained accuracy with different rival methods are reported in Table 1 using the UCF101 dataset. In addition, we have presented the accuracy of single classifiers, each trained on five different individual feature sets.

As can be seen, the ensemble-based approaches have remarkably improved the results. Specially, our third approach outperforms other state-of-the-art methods with an overall accuracy of 75.05% by averaging over the three training/testing splits. This is slightly better than [23, 18]

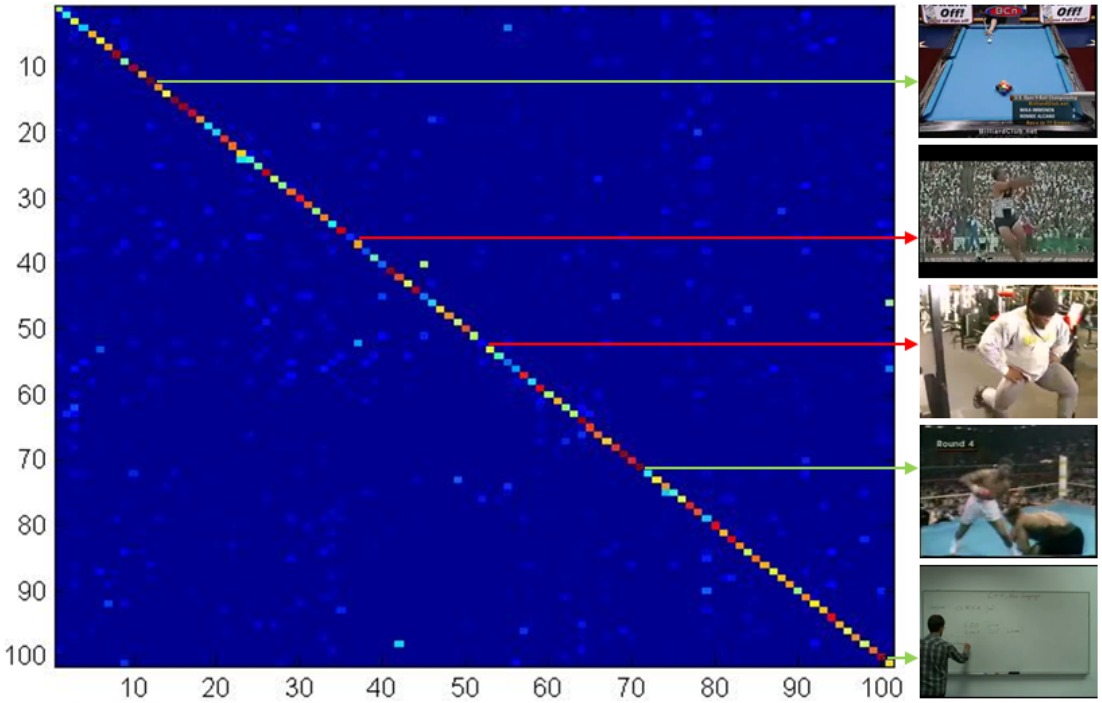

Figure 3. Confusion matrix of the ensemble classification system (third approach) for the UCF101 dataset. The green and red arrows point towards image examples of action classes of low (*billiards shot*, *punch* and *writing on board*) and high (*hammer throw* and *lunges*) confusion, respectively.

who reported an average accuracies of 73.39% and 73.10%, and remarkably better than work of Karpathy et al., which is based on Convolutional Neural Networks (CNNs), presented at CVPR 2014 [10]. In addition, the confusion matrix of the third approach for the UCF101 dataset is shown in Figure 3. In the figure, image examples of action classes of low and high confusion are given. In general, the actions which result in the highest amount of confusions, and thereby the lowest recognition accuracies, are actions videos affected by a high amount of camera and/or background motion.

Figure 4 shows the classification accuracy of the third ensemble-based approach as a function of the ensemble size for UCF101 datasets. These observation is consistent with the results of many studies, see [21, 9] as few examples, that is, the ensemble classification performance first improves as the ensemble size increases and then plateaus after a demarcation point, e.g., a value around 40-50 % accuracy.

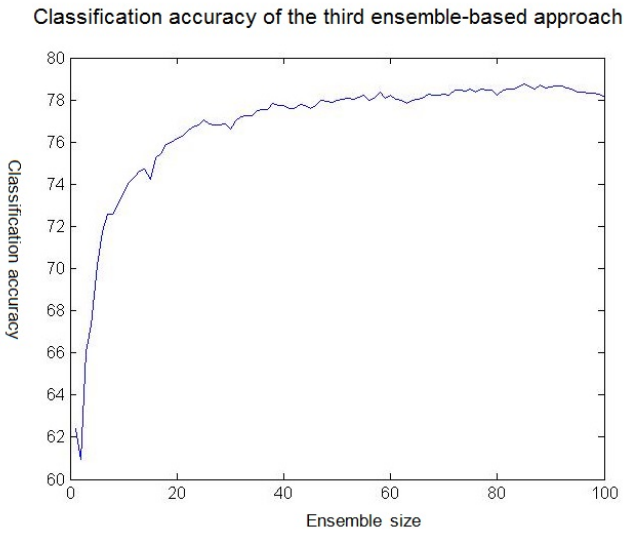

Figure 4. Accuracy of ensemble classification method versus the ensemble size.

5. Conclusion

This paper presents an ensemble classification framework to address the multiple action recognition problem. We designed a set of classifiers, each one trained over different feature sets. 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 UCF101 dataset, showing significant performance improvements of the proposed methodology. We also show performance improvements in relation

Table 1. Average recognition accuracies (%) for the UCF101 dataset in comparison to state-of-the-art.

Method (year)	Overall Acc.	Split 1 Acc.	Split 2 Acc.	Split 3 Acc.
Karpathy et al. [10] (2014)	65.4	-	-	-
Phan et al. [23] (2013)	73.39	71.10	73.67	**75.39**
Murthy et al. [18] (2013)	73.10	-	-	-
Rostamzadeh et al. [26] (2013)	70.50	70.45	69.80	71.27
Nga et al. [19] (2013)	66.26	65.16	66.73	66.90
Cho et al. [4] (2013)	65.95	65.22	65.39	67.24
Paez et al. [22] (2013)	65.68	65.31	65.48	66.23
Chen et al. [3] (2013)	64.30	63.41	65.37	64.12
Burghouts et al. [2] (2013)	63.46	62.01	63.46	64.90
Nga et al. [20] (2013)	60.10	-	-	-
Wang et al. [28] (2013)	54.74	54.76	55.16	54.29
Soomro et al. [27] (2012)	43.90	-	-	-
Single feature set				
STIP + BoVW	42.56	42.12	41.89	43.67
Dense Trajectory - HOF	51.10	50.19	51.76	51.35
Dense Trajectory - HOG	46.59	46.47	46.69	46.60
Dense Trajectory - MBH	62.93	62.54	62.78	63.46
Dense Trajectory - TR	49.88	49.76	50.05	49.83
Employing different feature set				
Approach 1: Baseline (Early feature fusion)	60.73	61.13	60.11	60.95
Approach 2: Ensemble-based	69.10	69.43	68.09	69.79
Approach 3: Ensemble-based (RSM)	**75.05**	**75.11**	**74.80**	75.23

to state of the art methods on the UCF101 dataset.

References

[1] X. Baro, J. Gonzalez, J. Fabian, M. Bautista, M. Oliu, E. H, I. Guyon, and S. Escalera. Chalearn looking at people 2015 challenges: action spotting and cultural event. In *CVPRW on ChaLearn Looking at People workshop*, 2015. 1

[2] G. Burghouts, P. Eendebak, H. Bouma, and R.-M. ten Hove. Action recognition by layout, selective sampling and soft-assignment. In *THUMOS: ICCV Workshop on Action Recognition with a Large Number of Classes*, 2013. 7

[3] W. Chen, R. Xu, and J. Corso. Action bank for large-scale action classification. In *THUMOS: ICCV Workshop on Action Recognition with a Large Number of Classes*, 2013. 7

[4] H. Cho, H. Lee, and Z. Jiang. Evaluation of LC-KSVD on UCF101 action dataset. In *THUMOS: ICCV Workshop on Action Recognition with a Large Number of Classes*, 2013. 7

[5] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *ECCV*, 2006. 2

[6] A. Dempster. Upper and lower probabilities induced by multivalued mappings. *Annals of Mathematical Statistics*, 38(2):325339, 1967. 3

[7] 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*, pages 65–72. IEEE, 2005. 2

[8] A. Hernandez-Vela, M. A. Bautista, X. Perez-Sala, V. Ponce, X. Baro, O. Pujol, C. Angulo, and S. Escalera. Bovdw: Bag-of-visual-and-depth-words for gesture recognition. In *ICPR*, pages 449–452. IEEE, 2012. 2

[9] T. K. Ho. The random subspace method for constructing decision forests. *TPAMI*, 20:832–844, 1998. 4, 6

[10] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014. 6, 7

[11] J. Kittler, M. Hatef, R. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Inteligence*, 20(3):226239, 1998. 3

[12] A. Kläser, M. Marszalek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC*, 2008. 2

[13] L. I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley, New York, NY, 2004. 2, 3

[14] I. Laptev. On space-time interest points. *IJCV*, 64(2-3):107–123, 2005. 2

[15] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008. 2

[16] J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *CVPR*, pages 3337–3344. IEEE, 2011. 2

[17] S. Maji, A. C. Berg, and J. Malik. Classification using intersection kernel support vector machines is efficient. In *CVPR*, pages 1–8. IEEE, 2008. 5

[18] O. Murthy and R. Goecke. Ordered trajectories for large scale human action recognition. In *CVPR*, 2013. 5, 7

[19] D. Nga, Y. Kawano, and K. Yanai. Fusion of dense SURF triangulation features and dense trajectory based features. In *THUMOS: ICCV Workshop on Action Recognition with a Large Number of Classes*, 2013. 7

[20] D. Nga and K. Yanai. A spatio-temporal feature based on triangulation of dense surf. In *THUMOS: ICCV Workshop on Action Recognition with a Large Number of Classes*, 2013. 7

[21] D. Opitz and R. Maclin. Popular ensemble methods: an empirical study. *J. Artif. Res.*, 11:169–198, 1999. 6

[22] F. Paez, J. Vanegas, and F. Gonzalez. Mindlab at the thumos challenge. In *THUMOS: ICCV Workshop on Action Recognition with a Large Number of Classes*, 2013. 7

[23] S. Phan, D.-D. Le, and S. Satoh. Nii, japan at the first thumos workshop 2013. In *THUMOS: ICCV Workshop on Action Recognition with a Large Number of Classes*, 2013. 5, 7

[24] R. Polikar. Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, 6(3):21–45, 2006. 2

[25] G. Rogova. Combining the results of several neural network classifiers. *Neural Networks*, 7:777781, 1994. 3

[26] N. Rostamzadeh, J. Uijlings, and N. Sebe. Action recognition using accelerated local descriptors and temporal variation. In *THUMOS: ICCV Workshop on Action Recognition with a Large Number of Classes*, 2013. 7

[27] K. Soomro, A. Zamir, and M. Shah. UCF101: A dataset of 101 human action classes from videos in the wild. In *CRCV-TR-12-01*, 2012. 4, 5, 7

[28] F. Wang, X. Li, and W. Shu. Experimenting motion relativity for action recognition with a large number of classes. In *THUMOS: ICCV Workshop on Action Recognition with a Large Number of Classes*, 2013. 7

[29] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *CVPR*, pages 3169–3176. IEEE, 2011. 2, 4

[30] H. Wang and C. Schmid. Action recognition with improved trajectories. In *ICCV*, 2013. 2, 4

[31] H. Wang, M. Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2009. 2

[32] G. Willems, T. Tuytelaars, and L. Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *ECCV*, pages 650–663. Springer, 2008. 2

[33] T. Windeatt. Accuracy/diversity and ensemble mlp classifier design. *IEEE Transactions on Neural Networks*, 17(5):1194–1211, 2006. 2