2016 13th Conference on Computer and Robot Vision

Human Action Recognition by Fusing the Outputs of Individual Classifiers

E. Mohammadi, Q.M. Jonathan Wu, M. Saif
Department of Electrical and Computer Engineering
University of Windsor, 401 Sunset Avenue, Windsor, N9B 3P4, Canada
Email: (moham12b, jwu, msaif)@uwindsor.ca

***Abstract*—the combination of outputs of single classifiers for human action recognition is proposed and evaluated in this paper. Combining the outputs of individual classifiers leads to having a more accurate and robust-applicable framework by reducing the risk of a weak choice of a classifier or a set of features. The weakness of single classifiers becomes more bold when the problem difficulty increases, particularly while having numerous action types or similarity of actions. In this paper, the individual support vector machines (SVMs) are trained using diverse feature sets from different perspectives. Then, the outputs of single SVMs are fused by employing the algebraic combinations and Dempster Shafer fusion methods. The experimental results show that the action recognition accuracy is improved while employing the algebraic combinations to fuse the outputs of single classifiers.**

***Keywords*—Action Recognition, Ensemble of Classifiers, Dempster Shafer Fusion, Algebraic Combinations of Single Classifiers**

I. Introduction

Human action recognition is considered as one of the active research topics due to its variety of real-world applications in video surveillance, gaming, health care, human computer interaction, and video retrieval. Although gesture and action recognition is a growing research area in computer vision, it is still a challenging problem due to the intra-class variations which are derived by factors such as the style and duration of a specified action. In addition to the intra-class variations, motion clutter caused by moving background objects, variability due to camera motion, and computational challenges due to the huge amount of data are considered as major action recognition challenges.

Activity recognition is a multi-class classification problem where each target class is assigned to a specific action type. An action classification system consists of two main stages: first, the features are selected, described, and encoded. Then, the extracted feature are fed to a learning model to recognize actions. In such a system, an efficient feature set can reduce the burden of the classification algorithm. However, a powerful classification algorithm is able to classify actions with high accuracy even with a low discriminative feature set.

Computer vision scientists have been working to evaluate different visual descriptions and representations for video which are well-suited for action recognition problems. The performances of different representations are typically evaluated on benchmark datasets [4]. Different feature sets are fed to classifiers to identify the superiority of some visual descriptions and representations over the others [4]. Histogram of Oriented Gradients (HOG), Histograms of Optical Flow (HOF), Motion Boundary Histograms (MBH), and trajectory approaches are the low-level features descriptors which attain remarkable results on different datasets compared to the other state-of-the-art descriptors [4]. Based on the recent research presented in [4], the Fisher Vector encoding is considered as one the most efficient encoding approaches for action recognition.

Generally, the power of several representation or description techniques can be modified by employing the following methods: The first method is to merge the extracted feature sets, which is called early fusion, and then to feed this higher dimensional feature set to an individual classifier. The second method is to train several individual classifiers, each trained on a separated feature set, and then efficiently fuse the outputs of single classifiers in a late fusion model. It should be noted that the discriminative power of encoded feature sets can not be completely utilized while employing individual classifiers. In other words, while dealing with difficult action recognition problems, mainly when working with many action types and/ or having resemblance between actions, the single classifiers are not able to accurately categorize actions. Therefore, an ensemble classification framework must be employed to boost the recognition performance, where each combination of a feature set and a single classifier is a human action learner [2]. As stated in [3], a strategic combination of the learners can significantly enhance the classification accuracy. In [3], the Dempster-Shafer and algebraic fusion methods are employed to fuse the outputs of several single classifiers. The joint efficiency of the ensemble of multiple classifiers can compensate for a deficiency in one learner while employing the fusion over several single classifiers [3].

The ensemble of classifiers has been presented to recognize actions using the dense trajectory features and bag of words encoding algorithm in [2]. However, we have utilized the improved dense trajectory and Fisher Vector encoding to extract and encode feature sets as presented in [1]. Then, the encoded feature sets are fed to the individual SVMs separately, and finally the outputs of single SVMs are fused to recognize actions. In addition to the DS fusion method which has been used in [2], product, mean, and maximum rules from the algebraic fusion techniques are implemented and evaluated in

978-1-5090-2491-9/16 $31.00 © 2016 IEEE
DOI 10.1109/CRV.2016.46

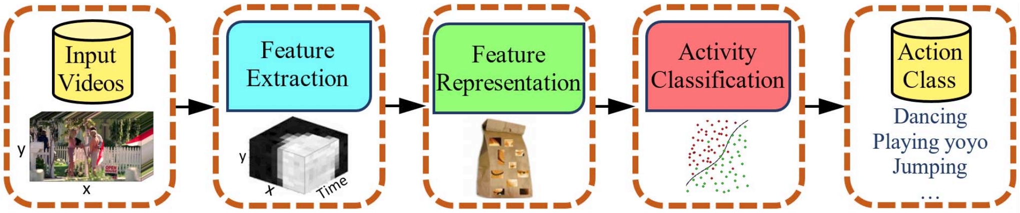

Fig. 1. The general action recognition framework

this paper.

The rest of the paper is organized as follows. The problem statements and related works about action recognition systems are briefly explained in Section 2. Section 3 describes the proposed framework to boost the action recognition performance. The experimental setup and results are demonstrated and evaluated in Section 4. Finally, the conclusion is presented in Section5.

II. Related Work

The general action recognition framework typically consists of feature extraction, feature representation, and classification phases as shown in Fig. 1. It should be noted that the key problem for successful action recognition is how to extract informative and robust features from the temporal motion and spatial gradient information. Numerous algorithms have been presented for feature extraction from videos. Local feature detection is considered as one of the most efficient approaches for action recognition. The main advantage of the local feature based methods is that no information on human body model or localization of people is required [4]. Local features are extracted by employing a local feature detector and then by encoding spatio-temporal neighbourhoods around the detected features using a local feature descriptor [5]. Local feature detectors for videos are divided into two following categories: spatio-temporal interest point (STIP) detector [5] and trajectory detector [6].

Action recognition with dense trajectories is presented in [6] and has been improved recently [1]. The motivation of the improvement was to remove spurious trajectories derived from camera motion in realistic videos. In the improved version, following the robust achievement of dense sampling over sparse sampling methods in the image domain, points are sampled densely from a multi scale pyramid by means of analysing each frame of a given video [1]. Then, the tracking of the detected points is performed for a certain time window. The dense optical flow field computation [7] is employed for tracking. It should be noted that the stated dense optical flow is applied on each spatial scale individually. Using a homograph matrix between sequential frames, it is possible to compensate for the camera motion to some degree [1]. Once the camera motion is known, the optical flow field can be re-computed to correct for such motion. In the next step, local descriptors, which are aligned with the trajectories, are extracted to encode local motion information. For each trajectory, 96-dimensional HOG, 108-dimensional HOF, and 192-dimensional MBH for x and y axis, and 30-dimensional trajectory features were extracted. All these descriptors are based on one-dimensional histogram of individual features. With this improvement, a very robust and accurate descriptors' performance is achieved in benchmark action recognition datasets [1].

The local descriptors that are derived from the space-time volume must be represented by a fixed size vector. Traditional methods learn a codebook from the training descriptors, and employ this codebook as a visual vocabulary to encode the local descriptors. The Bag of features (BOF) method has been widely used and adopted as the main model for video representation by pooling the local descriptors [8]. The general framework for traditional BOF contains local feature extraction, visual dictionary production with a clustering algorithm such as k-means and feature encoding. A better encoding approach, namely Fisher Vector (FV) representation, has been presented recently and increased the accuracy of recognition performance [9]. It should be noted that the FV approach does not represent features as a combination of visual words, but instead it represents differences between features and visual words. FV creates a visual vocabulary by clustering local features, extracted from the training videos, where clustering is done with Gaussian Mixture Model (GMM). Then, it captures the average first and second order differences between local features and visual vocabularies [9].

The classification is the last stage in the action recognition framework, where the SVM approach is a common classifier. SVMs maximize the distance between a hyperplane that divides two classes of data and instances on either side of it. Linear and non-linear separations can be performed using a kernel function in a SVM classifier. Moreover, they reach the global minimum and avoid ending in a local minimum [10]. According to the authors of [11], SVMs achieve the best accuracy in general, in comparison with the Decision Trees, Neural Networks, Nave Bayes, k-NN, and Rule-learners. They are also at least as good as others in speed of classification, tolerance to irrelevant attributes, tolerance to redundant attributes, and tolerance to highly interdependent attributes. Therefore, the SVM is considered as one of the most popular classifiers for action recognition. However, using a single classifier to classify patterns has been recently challenged by multiple classifier approaches, where the classification system is derived from an ensemble of single classifiers whose outputs are pooled in some way to attain a more accurate final classification decision. In an ensemble classification approach, each single classifier will focus on diverse aspects of the data and will err under different situations. It should be noted that single classifiers errors have to be uncorrelated in an ensemble classification system [3]. Consequently, the total errors can be reduced by an applicable combination of the single classifiers if each single classifier makes different errors.

An ensemble of SVMs, which is presented in [2], has been implemented for action recognition using the dense trajectory feature sets. As presented in [2], the BOF representation approach has been performed to encode the descriptors. Then, several single classifiers have been trained on different feature sets, and then fused using the DS fusion algorithm. In our paper, the improved dense trajectory along with the fisher vector encoding are employed to create the feature sets. The reason of choosing the FV method is that the achieved recognition accuracy using the FV is significantly higher than BOF using single classifiers [4]. In the classification stage, aside from the DS fusion, the product, mean, and maximum rules from the algebraic combination methods are employed to fuse the outputs of several individual classifiers [15]. To the best of our knowledge, human action recognition using the improved dense trajectory, FV encoding, and ensemble of SVMs by means of the algebraic combinations has not been implemented and evaluated previously.

III. Proposed Action Recognition

The efficiency of pattern classification using an individual classifier has been recently challenged by multiple classifier systems [3]. The underlying idea behind the ensemble of classifiers is that instead of employing a very complex representation/ learning technique, the action categories can be classified using a set of relatively simple and diverse classifiers, each trained with different feature sets [2]. In an ensemble classification system, it is hoped that each individual classifier will focus on different aspects of the data and will err under different situations [3].

In this work, as shown in Fig. 2, the encoded data from the individual descriptors are fed to the single classifiers separately. Then, the outputs of single classifiers are fused to classify actions. A combination function is employed to merge the outputs of single classifiers. A variety of combination functions to fuse the single classifiers have been presented in [3]. In this paper, the Dempster-Shafer (DS) fusion, product, mean, and maximum rules from the algebraic combination methods have been used to combine the outputs of single classifiers. The stated fusion methods make an ensemble of classifiers by observing the outputs of single classifiers as a measure of evidence which is derived from the source that is generated from the training data. This section briefly describes the feature extraction, encoding, and fusion methods which are employed in this paper.

A. Feature Extraction

The Improved Dense Trajectory (iDT) method , which has been introduced in [1], is employed for feature extraction. In the iDT, each frame of the video is analysed to sample the points densely from a multi scale pyramid. Then, the sampled points are tracked for a given time window. It should be noted that the employed tracking is based on dense optical flow field computation [7] and is applied on each spatial scale separately. Finally, local descriptor extractions are performed to encode local motion information.

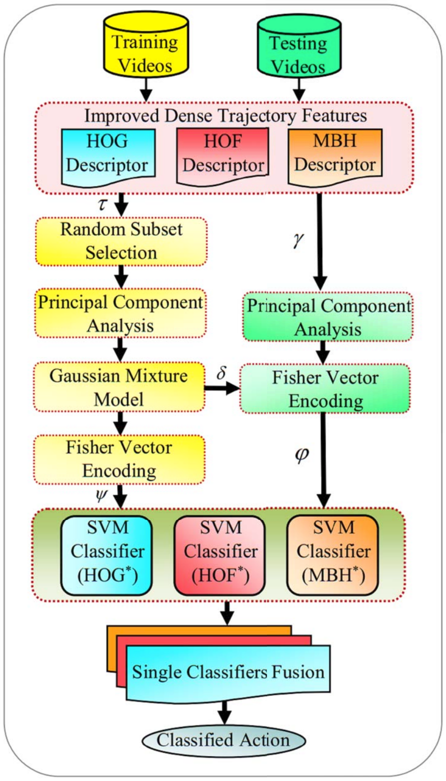

Fig. 2. Bold arrows stand for the three separated descriptors (HOG, HOF, and MBH); τ and γ respectively stand for the training and testing splits from the three separated descriptors; δ stands for the GMM vocabularies from the training data of three stated descriptors; ψ stands for the training feature sets encoded by the fisher vector; φ stands for the testing feature sets encoded by the fisher vector

For each trajectory, three descriptors (HOF, MBH, and HOG) are computed in the space-time volume with exactly the same parameters as stated in [6]. HOF and MBH are based on optical flow, and capture motion information. The orientation of flow vectors are quantized using the HOF descriptor. However, MBH divides the optical flow into horizontal and vertical components. Then, the derivatives of each component are computed by MBH descriptor. HOG is based on the orientation of image gradients, and measures the information from static appearance. The employed dimensions of the HOG, HOF, and MBH for x and y axis are 96, 108, 96 and 96 respectively as stated in [1].

B. Feature Encoding

The Fisher Vector (FV) approach is employed to encode the data from descriptors. FV has been reported to consistently boost the classification performance [16]. Another significant advantage of FV encoding is its good performance using fast linear classifiers. Encoding the visual vocabularies using FV has been proposed by [12]. FV requires Gaussian Mixture Models (GMMs) to shape the vocabulary, so that it encodes both first and second order statistics. The difference among pooled descriptors and the vocabulary is calculated by applying derivative operations on the likelihood with respect to the distribution parameters of the vocabulary. Even though MBH on x and y axis, HOF and HOG are all histogram based descriptors, they carry different information from different perspectives. Consequently, separate GMMs are learnt for each aspect. Four different Fisher vectors are produced using the four separated aspects. Each FV is normalized separately before feeding to single classifiers. The normalization is performed by sign square-rooting (power normalization) followed by L_2 normalization [12].

The GMM including K components is trained to learn the parameters $\lambda = \{\omega_k, \mu_k, \Sigma_k\}_{|_{k=1}^{k}}$ over a random subset of the training features. It should be noted that using all training descriptors is computationally expensive and requires large amounts of memory. Thus, the procedure of random subset selection is considered as a common practice in this case. We randomly select 1000 features from each descriptor. The random selection of features should not change the distribution we want to learn. Since the covariance matrices are considered to be diagonal, the PCA approach is applied prior to building vocabulary for decreasing the size of the resulting FV and to decorrelate features to support the diagonal covariance assumption. In this paper, K is considered to be equal to 128 and each descriptor is reduced to 64 dimensions. Given a video with a set of descriptors $\{x_1, ..., x_n\}$, the FV becomes the concatenation of the normalized partial derivatives of means and deviations, u_k and v_k as stated below:

$$u_k = \frac{1}{N\sqrt{\omega_k}} \sum_{i=1}^{N} \gamma_{ki} (\frac{x_i - \mu_k}{\Sigma_k}) \tag{1}$$

$$v_k = \frac{1}{N\sqrt{2\omega_k}} \sum_{i=1}^{N} \gamma_{ki} [(\frac{x_i - \mu_k}{\Sigma_k})^2 - 1] \tag{2}$$

where γ_{ki} is the posterior probability that signifies each vector with a component k in the GMM [12]. The final dimension of the data derived from the FV encoding approach becomes $2DK$, where D is the dimension of the descriptors.

C. Combining Ensemble Members

The single SVM classifiers are trained using different feature sets that are encoded by the Fisher Vector. Then, the outputs of single classifiers are fused to make the final decision for the test samples. A decision profile, as shown in Fig. 3, is employed to discover the overall support for each class, and then to label the test sample as the class with the largest support. It should be noted that the outputs of a single classifier for each class are interpreted as the degree of support given to that class, and under certain conditions are described as an estimate of the posterior probability for that class [15].

In a decision profile, each classifier D_i in the ensemble $\mathrm{E} = \{D_1, D_2, ..., D_L\}$ results c degrees of support considering $x \in \Re^n$ to be a feature vector and $\Omega = \{\omega_1, \omega_2, \omega_3, ..., \omega_c\}$ to be the set of class labels. All defined c degrees of support are considered to be in the interval [0, 1]. Classifier D_i defines that x comes from class ω_j by showing the $d_{i,j}(X)$ support. The class label ω_j is assigned to an instance, if the support of that class is the largest one compare to other supports. Fig. 3 shows the desicion profile $(DP(X))$ using the results of L classifiers for a particular instance X. The fusions of degrees of supports in a decision profile are performed by the Dempster Shafer and algebraic fusion methods which are briefly explained in this section.

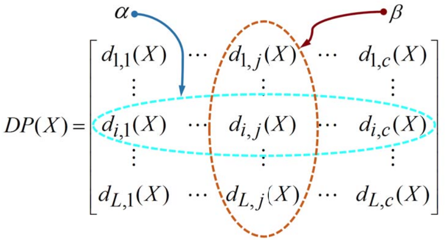
α
β
DP(X)=
d1,1(X) ⋯ d1,j(X) ⋯ d1,c(X)
⋮ ⋮ ⋮
di,1(X) ⋯ di,j(X) ⋯ di,c(X)
⋮ ⋮ ⋮
dL,1(X) ⋯ dL,j(X) ⋯ dL,c(X)

Fig. 3. α presents the output of classifier $D_i(X)$; β presents the support from classifires $D_1, ..., D_L$ for the class ω_j

1) Dempster Shapher Fusion: : The following four stages are performed to predict the target class of each test sample:

First, the decision templates are made by calculating the means of the decision profiles for all training samples belonging to the given class:

$$DT_j = \frac{1}{N_j} \sum_{z_k \in \omega_j} DP(z_k) \tag{3}$$

Second, the proximity is calculated between the decision templates and the output of classifiers as follows:

$$\phi_{j,i}(X) = \frac{(1 + ||DT_j^i - D_i(X)||^{-1}}{\sum_{k=1}^{c} (1 + ||DT_j^i - D_i(X)||^{-1}} \tag{4}$$

where DT_j^i refers to the i-th row of the decision template DT_j, and D_i denote the output of the i-th classifier (the i-th row of the decision profile $DP(X)$). The $\|.\|$ is a matrix norm in Eq. 4.

Third, the belief degrees are computed for each class and classification to show how a test sample is correctly assigned into a given class by a given classifier:

$$b_j(D_i(X)) = \frac{\phi_{j,i}(X)\Pi_{k \neq j}(1 - \phi_{k,i}(X))}{1 - \phi_{j,i}(X)[1 - \Pi_{k \neq j}(1 - \phi_{k,i}(X))]} \tag{5}$$

Finally, the Dempster rule is used to combine the belief degrees which are attained for each of the single classifiers. The Dempster rule states that the belief degrees from each classifier must be multiplied to achieve the final support for each class:

$$\mu_j(X) = K \prod_{i=1} b_j(D_i(X)), j = 1, ..., c \tag{6}$$

2) Algebraic Combiners: : The mean, maximum, and product rules are employed as the algebraic combiners to fuse the outputs of single classifiers [15]. In the pruduct rule, the total support for each class is calculated using a simple algebraic function of the supports received by single classifiers. As stated in Fig. 3, in the decision profile, each column represents the supports from classifiers and each row shows the outputs of each single classifier. The total obtained support for class ω_j is obtained from the column j of the decision profile as follows:

$$\mu_j(X) = F[d_{1,j}(X), ..., d_{i,j}(X), ..., d_{L,j}(X)] \tag{7}$$

where F is the following function called product combination rule:

$$\mu_j(X) = \frac{1}{L} \prod_{i=1}^{L} d_{i,j}(X) \tag{8}$$

The product rule selects the class whose product of supports from each classifier is the highest. In other words, the final decision is the class ω_j for which the total support $\mu_j(X)$ is the highest. It should be noted that the Product Rule decimates any class that obtains at least one zero or very small support due to the nulling nature of multiplying by zero.

For the mean rule, the support for a given class is the average of all classifiers' outputs for that given class as stated in Eq. 3. Thus, the F function is considered as an averaging function for the mean rule fusion. The final decision is the class ω_j for which the total support $\mu_j(X)$ is the highest. The maximum rule simply takes the maximum among the classifiers individual outputs where the ensemble decision is chosen as the class for which total support is largest.

D. Action Classification Approaches

Five approaches to recognize action categories are used and evaluated in this work. The first method is the baseline for comparing with the ensemble-based classifiers. As the first approach, the extracted feature sets are represented by Fisher Vector encoding and then merged. Consequently, a higher dimensional feature set is created by merging the individual feature sets. Then, this merged feature set is fed to a single SVM classifier to recognize action classes. As the second, third, fourth, and fifth methods, different feature sets are encoded using the FV encoding representation method. The employed feature sets are presented in the first column of table. 1. Then, each feature set is fed into its corresponding classifier. Finally, the outputs of these single classifiers are fused using the DS, and three algebraic combiners (product rule, mean rule, and maximum rule fusion methods).

IV. Experimental Setup and Results

In this paper, the issue of automatic recognition is addressed for human action recognition via supervised learning. It means that for every training video we know which action or actions it contains. The action models are learnt using training videos, and then these actions are recognized in new, unseen videos, i.e. videos for which we do not have annotations. This section briefly describes the procedure to implement the proposed method and describes the datasets which have been used to evaluate the recognition performance.

A. Datasets

The UCF 101, UCF Sports, and Weizmann datasets are used to evaluate the proposed method. UCF101 is a data set including 13320 videos and 101 action categories, consisting of realistic videos taken from YouTube [14]. UCF 101 is considered as a very challenging dataset due to large variations in camera motion, object appearance and pose, object scale, viewpoint, cluttered background, and illumination conditions [14]. The videos in 101 action categories are grouped into 25 sections, where each group consists of 4 to 7 videos of an action. It should be noted that the videos from the same group may share some common features, such as similar background and viewpoint. To divide all instances into train and test sets, the division procedure proposed by the authors of UCF101 is followed [14].

UCF Sports dataset consists of a set of actions from various sports which are typically featured on broadcast television channels [17, 18]. The dataset contains a total of 150 sample videos with the resolution of 720 by 480. The selection of videos represents a natural pool of actions with a variety of scenes and viewpoints.

The Weizmann Action Recognition dataset contains videos of 10 types of human actions [19]. Each action is implemented by 9 different people. Videos are captured with 180 by 144 pixels spatial resolution and 50 frames per second frame rate. In total, the dataset consists of 90 video sequences [19]. It should be noted that the low resolution videos, various people, and cloth variations are considered as the main challenges of the Weizmann dataset.

B. Classification

A set of visual feature sets are extracted using the most popular state-of-the-art descriptors for local features: Histogram of Oriented Gradients (HOG), Histogram of Optical Flow (HOF), and Motion Boundary Histograms (MBH) for x and y axis. The extracted feature sets are encoded using the Fisher Vector (FV) encoding method. The employed codebook size for the Gaussian Mixture Model in the FV encoding approach is 128. Then, the action learning models are trained by feeding the individual encoded feature sets to the single classifiers. For the classification, the SVM classifier with a cost equal to 100 and the linear kernel has been used as the single classifier. Next, the outputs of single classifiers are fused using the Dempster Shapher, product rule, mean rule, and maximum rule combiners. It should be noted that the individual classifiers are trained using different feature sets with a variety of dimensions in both feature space and sample length. As a result, it is shown that the derived predictions from single classifiers can be combined to boost the recognition performance. Following the multiple classifier philosophy, it is proofed that the proposed ensemble approach based on the

TABLE I
ACCURACIES OF ACTION RECOGNITION USING ENSEMBLE OF CLASSIFIERS

Feature Sets	Fusion Method	UCF101 Dataset	UCF Sports Dataset	Weizmann Dataset
HOG, HOF, MBH-x, and MBH-y	DS Fusion	84.42%	83.14%	**96.77%**
HOF and Merged HOG-MBH	DS Fusion	83.88%	81.71%	95.70%
HOG and Merged HOF-MBH	DS Fusion	82.73%	81.71%	93.55%
HOF, HOG, MBH	DS Fusion	81.74%	83.14%	96.77%
HOG, HOF, MBH-x, and MBH-y	Mean Rule Fusion	**86.21%**	**84.57%**	**96.77%**
HOF and Merged HOG-MBH	Mean Rule Fusion	85.72%	82.43%	95.70%
HOG and Merged HOF-MBH	Mean Rule Fusion	84.92%	82.43%	91.40%
HOF, HOG, MBH	Mean Rule Fusion	85.06%	81.71%	96.77%
HOG, HOF, MBH-x, and MBH-y	Maximum Rule Fusion	84.92%	81.00%	93.55%
HOF and Merged HOG-MBH	Maximum Rule Fusion	85.16%	82.43%	95.70%
HOG and Merged HOF-MBH	Maximum Rule Fusion	84.42%	81.71%	92.47%
HOF, HOG, MBH	Maximum Rule Fusion	84.02%	81.71%	94.62%
HOG, HOF, MBH-x, and MBH-y	Product Rule Fusion	83.88%	83.86%	**96.77%**
HOG and Merged HOF-MBH	Product Rule Fusion	85.76%	81.71%	91.40%
HOF, HOG, MBH	Product Rule Fusion	86.16%	81.71%	95.70%
HOF and Merged HOG-MBH	Product Rule Fusion	**86.21%**	81.71%	95.70%

product and mean rules outperforms standard non-ensemble strategies and other fusion methods for action recognition.

It bears mentioning that the Weizmann and UCF Sports datasets are evaluated using the Leave-One-Out cross-validation scheme. This scheme takes out one sample video for testing, and trains using all of the remaining videos of an action class. This is implemented for all the sample videos in a cyclic manner, and the overall accuracy is calculated by averaging the accuracy of all iterations. However, for the UCF 101 dataset, the proposed evaluation method in [14] has been employed to calculate the recognition performance.

C. Fusion Models

Several models of single classifiers have been trained using different feature sets to create the ensemble of classifiers. First, four seperated feature sets (MBH on x and y axis, HOG, and HOF) are trained by single SVMs and then fused using the DS, product, mean, and maximum fusion methods. Secondly, the MBH features on x and y axis were merged to yield a single MBH feature set. Thus, HOG, HOF, and MBH feature sets are trained using single SVMs and then fused in the next step. Third, HOF and MBH were merged to generate a single feature set. Then, the single SVMs were trained by the new merged HOF-MBH feature set and individual HOG. Fourth, The HOG and MBH feature sets were merged and with single HOF feature set have been used to train two single classifiers. Finally, all the feature sets are merged and then a single SVM classifier has been employed to train the merged feature sets.

D. Results

The attained recognition accuracies using Dempster Shafer, mean, maximum, and product fusion methods on the UCF101, UCF Sports, and Weizmann datasets are presented in table 1. In addition to the fusion methods, the accuracies of individual classifiers, each trained on separated feature sets, are shown in table 2. For each dataset, the highest achieved accuracy using specific feature sets is bold in both tables. The highest accuracy for the ensemble approaches are attained by training four single SVM classifiers based on HOF, HOG , and MBH for x and y axis descriptors, and fusion the outputs of these single classifiers using the mean rule approach. As shown in tables 1 and 2, the accuracy of ensemble of classifiers using mean rule fusion is higher than the accuracy of single classifiers even though using a merged feature set (early fusion) to train a single classifier. It must be noted that the score level fusion based on the product rule combination has improved the results for the UCF101 and Weizmann datasets due to the decimation of any class that obtains at least one zero or very small support respect to the nulling nature of multiplying by zero.

TABLE II
ACCURACIES OF ACTION RECOGNITION USING SINGLE CLASSIFIERS

Feature Set	UCF101	UCF Sports	Weizmann
HOG	74.84%	76.00%	84.95%
HOF	78.88%	80.29%	92.47%
MBH on x axis	77.82%	75.29%	93.55%
MBH on y axisy	78.15%	70.29%	78.49%
Merged of MBH-x and MBH-y	78.93%	79.57%	88.17%
Early Fusion of All Features	**85.06%**	**83.86%**	**92.47%**

V. CONCLUSION AND FUTURE WORKS

In this paper, the issue of automatic recognition is addressed for human action recognition via supervised learning. The efficiency of human action recognitions is enhanced by utilizing the Fisher Vector representation and improving the classification module. Each of the single classifiers are trained over different feature descriptors that are encoded by the Fisher Vector approach. The outputs of single classifiers are fused using the Dempster-Shafer and algebraic combiners to improve the performance of the action recognition. The classification performances that are derived from the single classifiers are compared with the results from fusing the classifiers. The experimental results show that the action recognition performance using the mean rule fusion is more accurate than the single classifiers, early fusion approach, and other employed fusion methods.

In the future work, linear discriminant analysis would be employed to decrease the dimension of the features. Extreme learning machine will be used in the classification stage to train a new model for action recognition. The extracted features using the deep-learning approaches and mid-level features also will be employed in the proposed method to modify the diversity of features in the ensemble of classifiers.

REFERENCES

[1] H. Wang, D. Oneata, J. Verbeek, C. Schmid. “A Robust and Efficient Video Representation for Action Recognition”, International Journal of Computer Vision, vol. 115, pp. 1-20, 2015.

[2] M. Bagheri, Q. Gao, S. Escalera, A. Clapes, K. Nasrollahi, M.B. Holte, T.B. Moeslund. “Keep it Accurate and Diverse: Enhancing Action Recognition Performance by Ensemble Learning”, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2015, pp. 22-29.

[3] L.I. Kuncheva. “Combining Pattern Classifiers: Methods and Algorithms”, Wiley, New York, 2004.

[4] P.T. Bilinski. “Human Action Recognition in Videos”, University of the Nice-Sophia Antipolis, 2014.

[5] I. Laptev, B. Caputo, C. Schuldt, T. Lindeberg. “Local velocity-adapted motion events for spatio-temporal recognition”, Computer Vision and Image Understanding, vol.108, 2007, pp. 207-229.

[6] H. Wang, A. Klaser, C. Schmid, C. Liu. “Dense trajectories and motion boundary descriptors for action recognition”, International Journal of Computer Vision, Springer , vol. 103 (1), 2013, pp.60-79.

[7] G. Farnebck. “Two-frame motion estimation based on polynomial expansion”, SCIA’03 Proceedings of the 13th Scandinavian conference on Image analysis, 2003, pp. 363-370.

[8] S. Ma, C. Zhang, J Sclaroff. “Action recognition and localization by hierarchical space-time segments”, IEEE International Conference on Computer Vision (ICCV), 2013.

[9] D. Oneata , J. Verbeek, C. Schmid. “Action and event recognition with Fisher vectors on a compact feature set”, IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1817-1824.

[10] L. Ladick, P. H.S. Torr. “Locally Linear Support Vector Machines”, 28th International Conference on Machine Learning, Bellevue, WA, USA, 2011.

[11] S. B. Kotsiantis. “Supervised Machine Learning: A Review of Classification Techniques”, Informatica J, vol. 31, 2007, pp.249-268.

[12] F. Perronnin, C. Dance. “Fisher kernels on visual vocabularies for image categorization”, IEEE computer society conference on Computer Vision and Pattern Recognition (CVPR), 2007, pp. 18.

[13] R. Messing, C. Pal, H. Kautz. “Activity recognition using the velocity histories of tracked keypoints”, IEEE International Conference on Computer Vision (ICCV), 2009.

[14] K. Soomro, A. Roshan Zamir and M. Shah. “UCF101: A Dataset of 101 Human Action Classes From Videos in The Wild”, CRCV-TR-12-01, November, 2012.

[15] C. Zhang, Y. Ma. “Ensemble Machine Learninig-Methods and Applications”, Springer, 2012, pp. 1-35.

[16] H. Jegou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, C. Schmid. “Aggregating local image descriptors into compact codes”, Transactions on Pattern Analysis and Machine Intelligence, 2012.

[17] M. D. Rodriguez, J. Ahmed, and M. Shah. “Action MACH: A Spatio-temporal Maximum Average Correlation Height Filter for Action Recognition”, Computer Vision and Pattern Recognition, 2008.

[18] K. Soomro and A. R. Zamir. “Action Recognition in Realistic Sports Videos”, Computer Vision in Sports. Springer International Publishing, 2014.

[19] L. Gorelick, M. Blank, E. Shechtman, M. Irani and R. Basri. “Actions as Space-Time Shapes”, Transactions on Pattern Analysis and Machine Intelligence, vol. 29(12), pp. 2247-2253, 2007.