Human Activity Recognition Using an Ensemble of Support Vector Machines

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*—numerous human action recognition algorithms have been developed and evaluated recently. However, the ensemble of classifiers to recognize actions, utilizing diverse feature sets, has remained untouched. Mixing the outputs of several classifiers decreases the risk of a weak choice of a learner or a set of features, and leads to having a more accurate and robust-applicable framework. The weakness of single classifiers becomes more evident when the problem difficulty increases, essentially while having numerous action types or resemblance of actions. In this paper, an ensemble of support vector machines (SVMs) is employed to improve the classification performance by fusing diverse features from different perspectives. The Dempster-Shafer fusion and product rule from the algebraic combiners have been utilized to combine the outputs of single classifiers. The experimental results show that the action recognition performance is improved while employing the ensemble of SVMs and stated fusion techniques.**

***Keywords*—Action Recognition, Dempster-Shafer Fusion, Product Rule Combiner**

I. Introduction

Nowadays, human action recognition is considered as an active research topic due to the wide applications in video surveillance, gaming, health care, human computer interaction, and video retrieval. Gesture and action recognition from a real-time visual stream is known as a challenging procedure for present computer vision approaches. Activity recognition is considered as a multi-class classification problem where each target class is assigned with a specific action type. An action classification system consists of two major steps: first, the features are selected, described, and encoded. Then, a classification algorithm is employed to categorize actions. In such a system, an efficient feature set is able to reduce the burden of the classification algorithm. On the other hand, a powerful classification algorithm is able to accurately classify actions even with a low discriminative feature set.

Computer vision scientists have been working to compare a variety of visual descriptions and representations for video which are well-suited for action recognition problems. The performances of these representations are typically evaluated on benchmark datasets [1]. Different feature combinations are fed to classifiers to identify the superiority of some visual descriptions and representations over the others [1]. Histogram of Oriented Gradients (HOG), Histograms of Optical Flow (HOF), Motion Boundary Histograms (MBH), and Trajectory approach are the low-level feature descriptors which attain remarkable results on different datasets compared to the other state-of-the-art descriptors [1]. Based on the recent research presented in [1], the Fisher Vector encoding is considered as one the most accurate encoding approaches for action recognition.

Generally, two methods are used to employ the power of several representation or description techniques. 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 single classifiers using separated feature sets, and efficiently fuse the outputs of the single classifiers in a late fusion model. It should be noted that the discriminative power of encoded information can not be completely utilized while employing individual recognition techniques. 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 recognition classifiers are not able to accurately categorize actions. Therefore, an ensemble classification framework can be employed to improve the recognition performance, where each combination of a feature set and a 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 Theory (DS) is 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 DS fusion over several single classifiers. Recently, the ensemble of classifiers has been presented to recognize actions using the dense trajectory features and the bag of words encoding algorithm [2]. However, we have utilized the improved dense trajectory and fisher vector encoding to select and encode the features. Then, the encoded features are fed to the same ensemble of SVMs which is presented in [2]. In addition to the DS fusion method, the Product Rule from the algebraic fusion techniques has been implemented and evaluated in this paper.

The rest of the paper is organized as follows. The problem statements and related works about action recognition systems are briefly described in Section 2. Section 3 describes the proposed framework for the action recognition problem. The experimental results are demonstrated and evaluated in Section 4. Finally, the conclusion is presented in Section 5.

978-1-5090-2088-1/16/$31.00 ©2016 IEEE

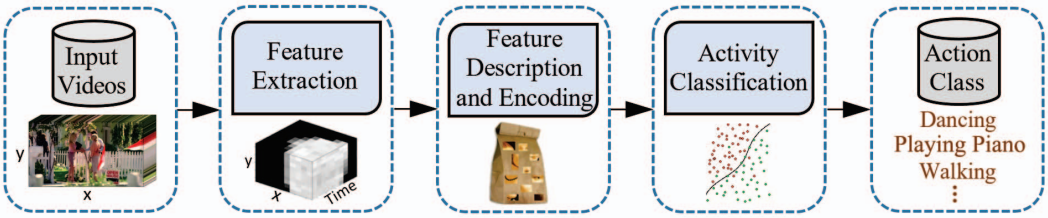

Fig. 1. The general framework for the human action recognition

II. Related Work and Problem Statement

The standard action recognition framework typically consists of feature sampling, description, encoding and classification phases as shown in Fig. 1. The key problem for successful action recognition is how to extract informative and robust features from the temporal motion and spatial gradient information. Numerous methods have been presented for detecting feature points from videos. Local features detection is considered as one of the most popular ways 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 [1]. 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 proposed in [6] and has been improved recently, as presented in [4]. The motivation of the improvement was to eliminate spurious trajectories created by camera motion in realistic videos and to advance the motion descriptors performance. 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 multiscale pyramid from each frame of the video [4]. Then, the detected points are tracked for a certain time window. The tracking is based on dense optical flow field computation [7] and is applied on each spatial scale individually. In the next step, local descriptors that 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 are extracted. All these descriptors are based on the one dimensional histogram representation of individual features, and they directly model values of the given features. With this improvement, a very robust and accurate descriptors' performance is achieved in benchmark action recognition datasets [4].

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 represent 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 enhanced 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. Firstly, it creates a visual vocabulary by clustering local features, extracted from the training videos, where clustering is done with Gaussian Mixture Model (GMM) clustering. Next, it captures the average first and second order differences between local features and visual vocabulary [9].

The classification is the last stage in a general action recognition framework, where the SVM approach is considered as the common employed classifier, with single or multiple kernel learning [10]. SVMs maximize the distance between a hyperplane that separates 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 are reliable in speed of classification, and tolerance to irrelevant, redundant, and 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 a way to attain a more accurate final 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. Next, the outputs of single classifiers are fused using the DS fusion algorithm. However, in this paper, aside from the DS fusion, the product rule fusion from the algebraic

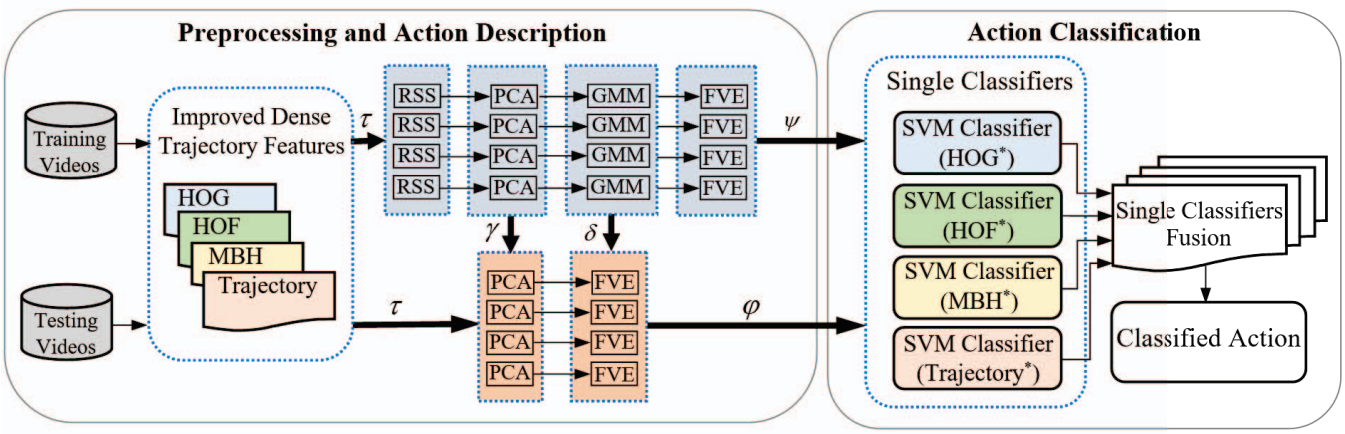

Fig. 2. The framework of the proposed activity recognition; RSS, PCA, GMM, and FVE respectively stand for Random Subset Selection, Principal Component Analysis, Gaussian Mixture Model, and Fisher Vector Encoding; τ stands for the four separated descriptors (HOG, HOF, MBH, and Trajectory); γ stands for the PCA projection coefficients; δ stands for the GMM vocabularies for four descriptors; ψ and φ respectively stand for the training and testing feature sets encoded by the fisher vector

combiners has been employed to fuse the outputs of individual classifiers. It should be noted that the FV representation algorithm is employed instead of the BOF to encode the descriptors which are based on the improved dense trajectory 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 [1]. To the best of our knowledge, human action recognition using the improved dense trajectory, FV encoding, and ensemble of SVMs has not been implemented and evaluated previously.

III. Proposed Action Recognition Framework

The efficiency of pattern classification by a single classifier has been recently challenged by multiple classifier systems [3]. The underlying idea of the ensemble of classifiers is that instead of employing a very sophisticated representation/learning technique, we can learn action categories using a set of relatively simple and diverse classifiers, each trained using different feature set [2]. 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 [3].

In this work, as shown in Fig. 2, the encoded data from the individual descriptors (HOG, HOF, MBH, and Trajectory) 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. The most common combination functions for fusing the single classifiers have been presented in [3]. In this paper, the Dempster-Shafer (DS) fusion method and the product rule fusion from the algebraic combiners have been used to combine the outputs of single classifiers. The stated fusion methods make an ensemble of classifiers by observing the output of single classifiers as a measure of evidence. This section briefly describes the fisher vector encoding, DS fusion method, and product rule fusion which have been employed for action recognition in this paper.

A. Fisher Vector Encoding

The fisher vector (FV) encoding is employed to represent the data from descriptors. Applying of the FV encoding on visual vocabularies has been recently proposed by [12]. The difference between the Bag of Features (BOF) and FV is that BOF requires k-means clustering and captures only 0_{th} order statistics, while FV requires gaussian mixture models (GMMs) to shape the vocabulary, so that it encodes both 1_{st} and 2_{nd} order statistics. FV encodes the difference among pooled descriptors and the vocabulary by applying derivative operations on the likelihood with respect to the distribution parameters of the vocabulary. The FV encoding provides a significant performance with a relatively few visual words, while it combines with simple linear classifiers. Even though Trajectory, 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 as shown in Fig. 2. Five different fisher vectors are produced using the five seperated aspects. All of the encoded feature vectors are normalized separately before feeding to single classifiers. The normalization is performed as in [12] by sign square-rooting (power normalization) followed by L_2 normalization.

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 of the training descriptors is computationally expensive and requires large amounts of memory. Thus, the procedure of random subset selection (RSS) is considered as a common practice in this case. The RSS procedure should not change the distribution we want to learn. The PCA approach is applied prior to building vocabulary for reducing the size of the resulting FV, and to decorrelate features to support the diagonal covariance assumption since the covariance matrices are considered to be diagonal. In this paper, K is considered to be equal to 64 to train the GMM and each descriptor is

reduced to 64 dimensions.

Given a video with the 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:

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

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

where γ_{ki} is the posterior probability that signifies each vector with a component K in the GMM [12]. The final dimension of the feature vector derived from the FV encoding approach becomes $2DK$, where D is the dimension of the descriptors. It is worth to mention that in the action classification's section of Fig. 2, the descriptors are shown by asterisk, which shows that the stated descriptors are encoded using the FV encoding method.

B. Dempster-Shafer Fusion

The single SVM classifiers are trained using the feature sets that are encoded by the fisher vector representation. Then, the outputs of single classifiers are fused to make the final decision about the test samples. The Dempster-Shafer theory (DS) is a mathematical theory of evidence based on belief functions and plausible reasoning, which is used to combine separate pieces of information to calculate the probability of an event. In the DS, the beliefs in a hypothesis are calculated as the sum of the masses of all sets it encloses. In other words, the DS allows combining evidence from different sources and arriving at a degree of belief (represented by a belief function) that takes into account all the available evidence.

As shown in Fig. 3, a decision profile is employed to discover the overall support for each class and then to label the test sample in the class with the largest 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, each classifier D_i in the ensemble $\mathrm{E} = \{D_1, D_2, ..., D_L\}$ results c degrees of support. 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 presents the desicion profile $(DP(x))$ from the L classifier outputs for a particular instance x.

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:

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

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

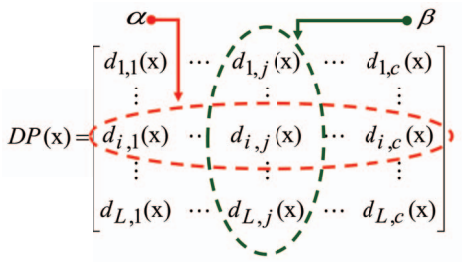

Fig. 3. α shows the output of classifiers $D_i(x)$; β shows the support from classifiers $D_1, ..., D_L$ for the class ω_j

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

where DT_j^i denotes the i_{th} row of the decision template DT_j, and D_i refers to 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:

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

Finally, the dempster rule is applied to combine the belief degrees which are derived from each of the single classifiers. Based on the dempster rule, the belief degrees from each classifier must be multiplied to achieve the final support for each class:

$$\mu_j(x) = \prod_{i=1} b_j(D_i(x)), j = 1, ..., c \quad (6)$$

C. Product Rule Fusion

Based on the pruduct rule, the total support for each class is calculated using a simple algebraic function of the supports that are received by single classifiers [15]. 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:

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

where F is the following function called product combination rule:

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

The product rule selects the class whose product of supports from each classifier 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.

D. Action Classification Approaches

Two classification models have been employed and evaluated in this paper. The first method is the baseline for comparing with the ensemble-based methods. As the first approach, the extracted feature sets are encoded by FV and merged. Thus, a higher dimensional feature set is made by combining the individual feature sets. Then, this higher dimentional feature set is fed to a single SVM classifier to recognize actions. As the second approach, each single feature set is encoded by FV, and fed to the single SVM classifier. Then, the outputs of these single classifiers are fused using the DS and product 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 describes the datasets and the experimental results for the proposed action recognition framework.

A. Datasets

The proposed action recognition approach is evaluated on URADL and UCF101 datasets. The URADL is a high resolution dataset of complicated actions, and contains 10 classes and 150 videos (15 videos for each class). The 10-fold cross validation method has been employed to test the classification performance during the training and testing of the URADL dataset.

UCF101 is another human action dataset including 13320 videos and 101 classes, consisting of realistic videos taken from YouTube. UCF 101 is considered as a very challenging task 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 was followed [14].

B. Experimental Results

Five different action descriptors (HOF, HOG, MBH on x and y axis and Trajectory), which are aligned with the dense trajectory features, are employed to extract features from the videos. These extracted features are then represented using the FV encoding method. The employed codebook size for the gaussian mixture model, in the FV encoding approach, is 64. The data derived from the FV encoding method are fed to the classifiers. In the classification stage, the SVM classifiers with a cost of 100 and the linear kernel have been used as the single classifiers. It should be noted that the action recogniton is performed using the following two models: First, as the early fusion method, the individual feature sets are merged to create a higher dimentional feature set. Then, the higher dimentional feature set is fed to a single SVM classifier to recognize actions. Second, as the late fusion method, the enemble of classifiers are trained using the five extracted feature sets. Five models of the feature sets to create the ensemble of classifiers are presented as follows:

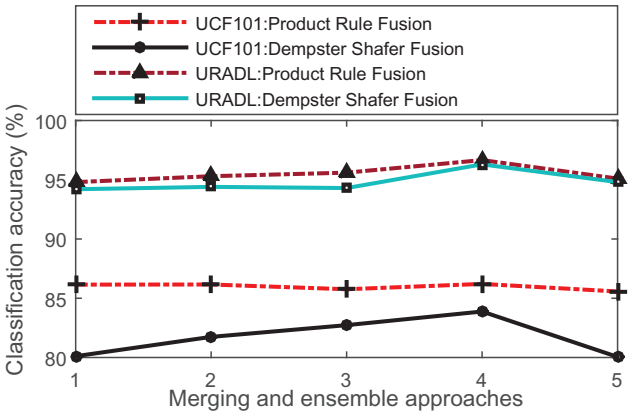

Fig. 4. The attained accuracies using five types of merging and ensemble approaches for UCF101 and URADL datasets

1) First model: five separated feature sets (Trajectory, HOG, HOF, MBH on x and y axis) are employed to train five single SVMs. Then, the outputs of single SVMs are fused using the DS and product rule fusion methods.

2) Second model: the Trajectory feature set was removed due to its inconsistency, and MBH features on x and y axis are merged to yield a single MBH feature set. Thus, HOG, HOF, and MBH feature sets are used to train three single SVMs. Then, the outputs of the three single SVMs are fused to build a late fusion model.

3) Third model: HOF and MBH are merged to generate a single feature set. Next, the single SVMs are trained by the new HOF-MBH feature set and HOG. Then, the outputs of the stated single classifiers are fused.

4) Fourth model: the HOG and MBH feature sets are merged. Then, the merged HOG-MBH and single HOF feature sets are used to train two single SVMs. Next, the outputs of these single classifiers are fused using the DS and product rule fusions.

5) Fifth model: the HOG and HOF feature sets are merged, and fed to first single SVM classifier. The second single SVM is trained using the MBH feature set. Finally, the outputs of two singlel classifiers are fused by DS and product rule fusions.

The accuracies of the above mentioned models for the URADL and UCF101 datasets are shown in Fig. 4. As shown in Fig. 4, the highest accuracies for the ensemble approaches are attained by training two single SVM classifiers based on HOF and merged of HOG-MBH descriptors, and fusion the outputs of these single classifiers using the product rule combiner. The accuracies of the single classifiers, early fusion approach, and the most accurate ensemble approach

TABLE I. Action Recognition Accuracies

Method	URADL	UCF101
Trajectory and Single SVM	78.00%	41.55%
HOG and Single SVM	83.33%	74.84
HOF and Single SVM	90.00%	78.88
MBH-x and Single SVM	87.33%	77.82%
MBH-y and Single SVM	88.67%	78.15%
Early Fusion and Single SVM	94.00%	85.06%
Ensemble-Based DS Fusion	96.30%	83.88%
Ensemble-Based Product Rule Combination	**96.66%**	**86.21%**

are presented in Table. 1. It must be noted that the score level fusion based on the product rule combination has improved the results on both 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.

V. Conclusion and Future Works

In this paper, the efficiency of human action recognitions is enhanced by improving the classification module and utilizing the fisher vector representation. A set of visual feature sets are extracted using the most popular state-of-the-art descriptors for local features: Histogram of Oriented Gradients, Histogram of Optical Flow, Motion Boundary Histograms for x and y axis, and Trajectory. The mentioned descriptors are encoded using the fisher vector encoding technique. Then, the action learning models are trained by feeding the individual encoded descriptors to single SVM classifiers. Next, the outputs of single classifiers are fused based on the Dempster-Shafer and product rule fusions. The classification performance of the single classifiers are compared with the results from fusing the classifiers. It is shown that the proposed ensemble approach based on the product rule outperforms standard non-ensemble strategies for action recognition.

In the future work, the extreme learning machine will be employed in the classification stage to boost the recognition performance. 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 for the ensemble of classifiers.

References

[1] P.T. Bilinski, "Human Action Recognition in Videos," Ph.D. dissertation, University of the Nice-Sophia Antipolis, 2014.

[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," CVPR Workshops, 2015, pp. 22-29.

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

[4] H. Wang, D. Oneata, J. Verbeek, C. Schmid, "A Robust and Efficient Video Representation for Action Recognition," IJCV, Springer, vol. 115, 2015, pp. 1-20.

[5] I. Laptev, B. Caputo, C. Schuldt, T. Lindeberg, "Local Velocity-Adapted Motion Events for Spatio-Temporal Recognition," CVIU, vol.108, 2007, pp. 207-229.

[6] H. Wang, A. Klaser, C. Schmid, C. Liu, "Dense Trajectories and Motion Boundary Descriptors for Action Recognition," IJCV, Springer , vol. 103 (1), 2013, pp.60-79.

[7] G. Farnebck, "Two-Frame Motion Estimation Based on Polynomial Expansion," SCIA, 2003, pp. 363-370.

[8] S. Ma, C. Zhang, J Sclaroff, "Action Recognition and Localization by Hierarchical Space-Time Segments," ICCV, 2013.

[9] D. Oneata , J. Verbeek, C. Schmid, "Action and Event Recognition with Fisher Vectors on a Compact Feature Set," ICCV, 2013, pp. 1817-1824.

[10] L. Ladick, P. H.S. Torr, "Locally Linear Support Vector Machines," ICML, 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," CVPR, 2007, pp. 18.

[13] R. Messing, C. Pal, H. Kautz, "Activity Recognition Using the Velocity Histories of Tracked Keypoints," 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, Nov, 2012.

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