Regression-Based Feature Selection on Large Scale Human Activity Recognition

Hussein Mazaar
Faculty of Computers & Info.
Cairo University, Egypt

Eid Emary
Faculty of Computers & Info.
Cairo University, Egypt

Hoda Onsi
Faculty of Computers & Info.
Cairo University, Egypt

***Abstract*—In this paper, we present an approach for regression-based feature selection in human activity recognition. Due to high dimensional features in human activity recognition, the model may have over-fitting and can't learn parameters well. Moreover, the features are redundant or irrelevant. The goal is to select important discriminating features to recognize the human activities in videos. R-Squared regression criterion can identify the best features based on the ability of a feature to explain the variations in the target class. The features are significantly reduced, nearly by 99.33%, resulting in better classification accuracy. Support Vector Machine with a linear kernel is used to classify the activities. The experiments are tested on UCF50 dataset. The results show that the proposed model significantly outperforms state-of-the-art methods.**

***Keywords*—Action Bank; Template Matching; SpatioTemporal Orientation Energy; Correlation; R-Squared; Support Vector Machine; Logistic Regression; Linear Regression; Human Activity Recognition**

I. Introduction

Human activity recognition is an active research area in artificial intelligence, human-computer interaction and computer vision. Applications of human activities include patient monitoring systems, surveillance systems, interfaces, virtual reality, motion analysis, robot navigation, robot recognition, video indexing, browsing, choreography,..etc. Human activities are conceptually partitioned based on their complexity into four different categories: gestures, actions or activities, group activities and interactions. Nowadays, digital cameras can record the most daily activities of people and this makes the video sources to be rich on the internet, and also brings the problem of video categorization and how a new input video is classified based on their activities classes. Generally speaking, the process of classification of input videos movies in the real world is impossible, also, the manual task is time-consuming. Many researchers engage a lot of attention to these problems. They tried to create a machine recognition model which the feature descriptors originated from the training videos are trained to automatically recognize the activities of the new videos [1], [2], [3].

Feature selection is a significant step in human activity recognition to identify the minimum number of features that improve the accuracy of the model. Moreover, the models with the smallest number of features can be simpler and faster in building and understanding. In general, the main types of feature selection are filters, wrappers, and embedded machine learning. The last type selects the features based on integration with machine learning.

Filters methods depend on the properties of the data to evaluate the features and are independent regarding learning methods, but they use statistical methods like information gain, correlation to calculate splitting criterion for decision tree. These statistical methods evaluate how well each feature partitions dataset. Wrapper methods measure the features based on the estimates or results of machine learning algorithms which integrate predictive estimates as feedback. One of the common methods is regularization, which uses in the optimization process of learning in predictive modeling as penalization. This approach penalizes the irrelevant features(coefficients) and selects the most important features to reduce the complexity (over-fitting) like LASSO, Ridge regressions. Feature selection in embedded methods performs in the training process of machine learning. It is efficient because no need for splitting data into training and validation sets. Also the approach is fast due to the re-training of a feature is not necessary. Wrapper methods provide better results than filters, but the computational cost is increased. Embedded methods have good results between performance and cost [4], [5].

The organization of this paper is structured as follows. In Section II, we discuss related work. Section III presents the Model framework. Section IV presents the feature detection based on spatiotemporal orientation energy and the detected features are described based on maximum pooling of template matching. Section V presents the feature selection process which mainly based on the R-squared regression model. Support vector machine is introduced in VI. Section VII shows the simulation results and the conclusion of the paper is summarized in section VIII.

II. Related Works

At the present time, local spatiotemporal features are the most public techniques of video representation. The techniques of local spatiotemporal features depend on detectors and descriptors. The detectors capture spatiotemporal interest point locations, like, Cuboids [6] and Harris3D [7]. The descriptors are extracted by HOG3D [8] or HOG/HOF [9]. Then pre-learned codebooks are defined to quantify the extracted features. Bag of Visual Words (BoVW) [10] can model videos. The local descriptors are local and repeatable features which are suitable advantages in video representation. They describe

appearance and motion information of a local cuboid nearly interest point. Due to simplicity and repeatability, the local descriptors are robust to deformation and intra-class variability. The drawback of local descriptors that They only display low level information, not high level motion, which makes the features lack discriminative power. Many recent researchers try to fix the issues by developing high level models like Silhouette [11], Space-time Shape [12], Motion Energy and History Image [13]. The recent approach is Actionbank [14]. A large combination of activity detectors are applied on input videos and the responses are used as rich representation for videos. The detectors are composed of global templates of activities which are discriminating and global. However, the global features are sensitive to deformation and intra-class variations.

III. The Proposed Framework

The proposed model of human action recognition is composed of four steps: feature detection, feature description, feature selection and classification (See Fig. 1). For each step, the algorithms are described in details in the following sections.

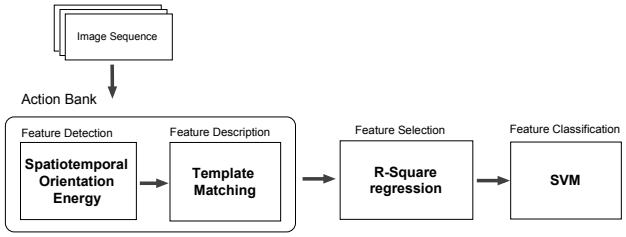


Fig. 1: The Proposed Framework of Human Action Recognition

IV. Feature Detection and Description

The videos are showed via high level features. The Action Bank [14] is the representation of videos. It is similarly close to object bank [15]. It represents the video as composed action detectors that each produces a correlation volume. The base element of feature is the template-based action detector. It is invariant/robust to variations in appearance, scale, viewpoint,and tempo.

A. Spatiotemporal Orientation Energy

Motion energies can represent an activity or video in various spationtemporal orientation. A composition of energies along various space-time orientations can capture the motion at a point during decomposition of video. These energies are the basis for low level activity representation. A decomposition of spatiotemporal orientation energies is performed using third derivatives of 3D Gaussian steerable filter which represents the strength of motion and used as local filter. Let $G^3_{\hat{\theta}}(x)$ denotes 3D Gaussian third derivatives, where $x = (x, y, t)$ indicates for location of spatiotemporal space and $\hat{\theta}$ denotes for unit vector of 3D directions. The spatiotemporal orientation energy is computed at every pixel as follows:

$$E_{\hat{\theta}}(x) = \sum_{x' \in \Omega(x)} (G^3_{\hat{\theta}} * V)^2 \quad (1)$$

where $\Omega(x)$ denotes for a local region around x, $V \equiv V(x)$ denotes for input video, and (*) indicates for convolution. Gaussian filters are separable filter that has some properties like estimation spatiotemporal orientation energy without executing convolution for all directions. The result of convolution is summed and squared over neighborhood space time Ω to get the energy measurement.

Marginalization for energy is a process to eliminate spatial orientation influence. Formally, the computation of energy with normal $\hat{n}$ at frequency domain plane $E_{\hat{\theta}_i}(\hat{n})$ by a simple sum

$$E'_{\hat{n}}(x) = \sum_{i=0}^{N} E_{\hat{\theta}_i \hat{n}}(x) \quad (2)$$

where N denotes for is Gaussian derivatives order, $\hat{\theta}_i$ is one of $N + 1 = 4$ directions calculated from Eqn. 2.

Officially $\hat{\theta}_i$ is provided by,

$$\hat{\theta}_i = cos\left(\frac{\pi i}{4}\right)\hat{\theta}_a(\hat{n}) + sin\left(\frac{\pi i}{4}\right)\hat{\theta}_b(\hat{n}), \quad (3)$$

where $\hat{\theta}_a(\hat{n}) = \hat{n} \times \hat{e}_x / \|\hat{n} \times \hat{e}_x\|$, $\hat{\theta}_b(\hat{n}) = \hat{n} \times \hat{\theta}_a(\hat{n})$,ê is the unit vector along the spatial x axis in the Fourier domain and $0 \leq i \leq 3$. The implementation for detectors of action bank,

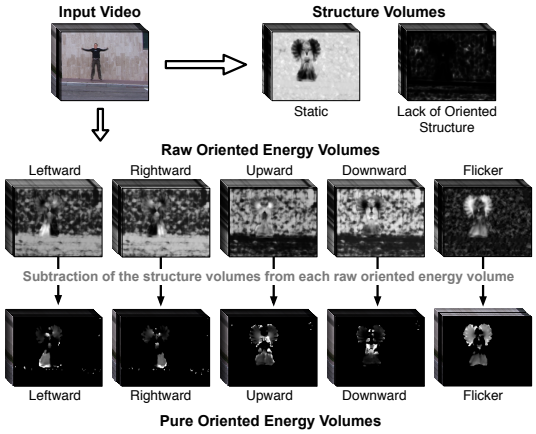


Fig. 2: A spatiotemporal orientation energy representation [14]

seven raw spatiotemporal energies are defined with different velocities:static E_s, leftward E_l, rightward E_r, upward E_u, downward E_d, flicker E_f , and lack of structure E_o. The lack of structure energy is calculated as function of six other energies and has peaks when no strong response from other six energies. The goal of this energy is to eliminate the instabilities of small energy points and gets a saliency. The pure energies are extracted from energies with subtraction of background and noise and are normalized to avoid influence of illumination adjustment and contrast as follows:

$$\hat{E}_i = max(E_i - E_o - E_s, 0), \ \forall i \in \{f, l, r, u, d\} \quad (4)$$

B. Template Matching

Detection an activity of small video called "template video" in a large video called "search video" is performed by scanning a 3D template video over all positions in spacetime. The similarity is determined by calculating each location among

histogram of oriented energy of the template and search video. The "action spotting" algorithm is the recent detector which is applied due to appropriate features of in-variance to activity localization, appearance variation, natural explanation like the decompose oriented energies and efficiency [16], [14]. The correlation between template video T and search video or query video is calculated by Bhattacharya coefficient m(.) as follows:

$$M(x) = \sum_{u} m(T(u), V(x-u)) \tag{5}$$

where M() denotes for the results of correlation and u denotes for ranges of template video. The correlation is efficiently performed in frequency domain and the output value is between 1 denoting full match or complete match and 0 denoting a complete mismatch which interprets volumetric max-pooling method.

Let N_a denotes for number of detectors for a given action bank and N_s denotes for scales of activity (run times), the output of correlation volumes are $N_a \times N_s$. The max-pooling technique in [17] is adapted as in Fig. 3 to be three levels in the octree which is $1^3+2^3+4^3$ or a 73 dimension vector [14]. For each activity, the total length of feature vector equals to $N_a \times N_s \times 73$.

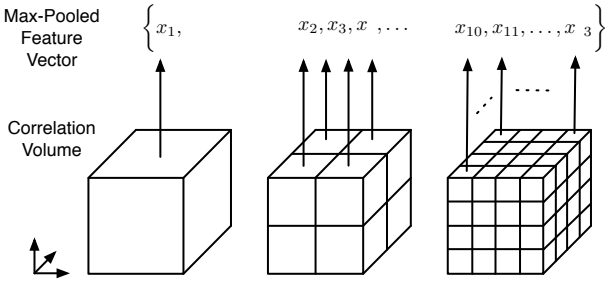


Fig. 3: Volumetric max-pooling technique [14]

V. Feature Selection

Feature selection is an important area in predictive modeling and statistics. Theory and practice of feature selection have shown that feature selection is an effective way in improving learning, enhancing recognition accuracy and decreasing complexity of human activity recognition. The objective of feature selection in supervised learning produces higher classification accuracy [18], [19], [20].

One of the most crucial issues in high-dimensional data is determining which features should be included in a model of human activity recognition. From a practical point of view, a model with less features may be more interpretative and less complexity. Statistically speaking, the model with less features is often more attractive. Also, some models are negatively affected by irrelevant features [21], [20].

R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression. It uses a forward stepwise least squares regression that maximizes the model R-squared value. In this model, the features assessment are fast provided as a preparatory step and the predictive models are rapidly simplified in development with huge data. Linear models can quickly identify input useful features for classifying the target classes. The R-Squared feature selection criterion has applied two steps processes as follow:

A. Squared Correlations

The squared correlation coefficient is the ratio of single input feature explains the variation in target class with elimination of other features in calculations. Also, It is called Coefficient of Determination (CoD) in statistics. The value ranges of squared correlation coefficient are between 0 (no relationship between the target class and input feature) and 1 (the variation of target class is totally explained with input feature). In human activity recognition, all input features are interval, so the squared correlation coefficient is calculated by a simple linear regression as follow:

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{6}$$

where Y denotes response variable or target, X denotes for input feature, β_0 denotes for intercept parameter, β_1 denotes for slope parameter and ε indicates the error deviation of Y about $\beta_0 + \beta_1 X$ (See Fig. 4a).

The feature has a significant influence if it explains the target, so the simple linear regression model is compared to the baseline model (Fig. 4b). The baseline regression has a horizontal fitted regression line over any value in input feature with slope equals to 0 and the intercept equals to the mean of response target $\bar{Y}$.

Explained variability is the distinction between the regression line and baseline line. The regression sum of squares (SSR) is the amount of variability explained by your model. The comparison between the explained variability to unexplained variability determines the amount of variability explained by regression line rather than baseline line. The Fig. 4c shows a seemingly contradictory relationship between explained, unexplained and total variability. The regression sum of squares (SSR) is equal to

$$\sum (\hat{Y}_i - \bar{Y})^2 \tag{7}$$

Unexplained variability is the distinction between the between the actual values and the regression line. The error sum of squares (SSE) is the amount of variability unexplained by regression model. The error sum of squares is equal to

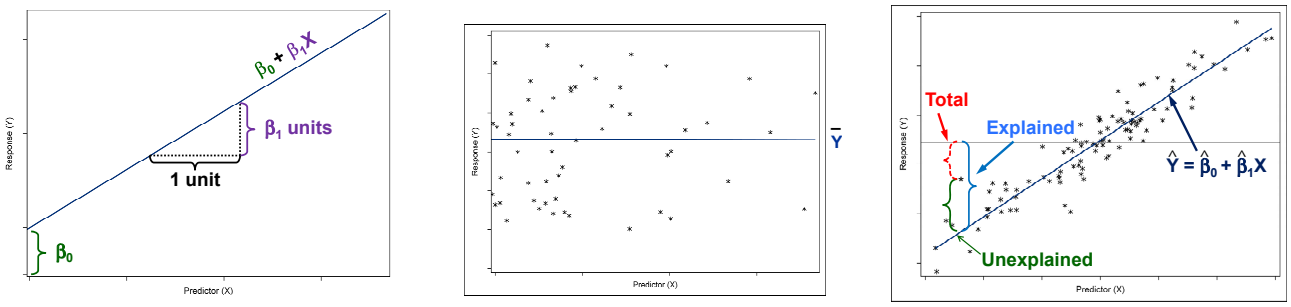
$$\sum (Y_i - \hat{Y}_i)^2 \tag{8}$$

Total variability is the distinction between the actual values and baseline regression line. The corrected total sum of squares (SST) is the sum of the explained and unexplained variability. The corrected total sum of squares is equal to

$$\sum (Y_i - \bar{Y})^2 \tag{9}$$

R-Squared the proportion of variability observed in the data explained by the regression line. The R-Squared is equal to

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \tag{10}$$



(a) Simple Linear Regression (b) Baseline Regression (c) Expained vs. Unexplained Variabilty

Fig. 4: Regression Model

B. Forward Stepwise Regression & Logistic Regression

This algorithm is applied after calculating the squared correlation coefficient for all input features in human activity recognition, the other important features are measured using a forward stepwise R-Squared regression. The sequential forward regression chooses The feature that has the highest squared correlation coefficient which explains the largest amount of variation in class target. At each iteration, the additional input feature is selected that gives the largest incremental increase in model of R-Squared. The stepwise algorithm ends when no other input feature can meet the Stop R-Squared criterion. The final logistic regression analysis is performed using the predicted values that are output from the forward stepwise selection as the independent input.

VI. Support Vector Machine

The concluding stage of the recognition process is the classification of the extracted features into a predefined set of classes. The field of machine learning has many powerful classification models. Our goal in this stage is to contribute to this field by introducing a reliable, accurate and interaction-centric classifier.

The human activities recognition are formulated by multi-calss classification problem. Each activity is represented by each class. The goal is assigning and classifying a video sequence to classes of activities. Many supervised learning methods are learned to activity recognizer. Support Vector Machine (SVM) is one of the superior machine learning in human activity recognition and high dimensional data because the prime generalization strength and highly accurate results. SVM can avoid over-fitting in neural networks based on risk minimization theory. Also, SVM can handle a high dimensional space by creating a maximal hyperplane to separate non-overlapping classes. Two parallel hyperplanes are proceeded in SVM and the goal of SVM seeks to find the maximal distance between the parallel byperplanes (Fig. 5). The better the classification, the larger the distance between byperplanes and vice vera.

Formally, Let the data set of training is $\mathbf{D} = \{\{x_i, y_i\}_{i=1}^{n} \mid x_i \in \Re^d, y_i \in \{-1, +1\}\}$ with n observations in a d-dimensional space and y_i denotes for classes, SVM can handle non-separable observation by slack variable ξ_i for observation x_i which indicates how much the observation violates the soft

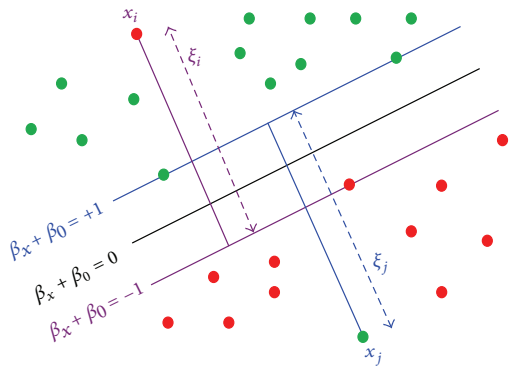


Fig. 5: Support Vector Machine with Slack Variables

margin constraints. The values of slack variable have three type: $\xi_i = 0$ denotes the observation away with at least $\frac{1}{\|W\|}$ from the hyperplane, $0 \leq \xi_i \leq 1$ denotes the observation between margins and when $\xi_i \geq 1$ then the observation is wrongly classified and appears on the wrong side. This approach achieves best performance for SVM. The quadratic programming can determine the optimal generalized separating hyperplane as follow:

$$\arg\min_{w,b,\xi_i} \frac{1}{2}\|W\|^2 + C\sum_{i=1}^{n}(\xi_i)^k \tag{11}$$

Subject to $y_i(w^T x_i + b) \geq 1 - \xi_i \wedge \xi_i \geq 0 \ \forall x_i \in \mathbf{D}$.

The parameter C is a constant called "regularization constant" to control the misclassification cost which governs the trade-off among maximal margins and minimal loss. The term $\sum_{i=1}^{n}(\xi_i)^k$ denotes for loss. The constant k controls the loss which becomes hinge loss when k is 1 and quadratic loss when k is 2. Dual formulation is recommended to solve SVM due to computational purposes. This solution uses Lagrangian method and is optimized with Lagrange multiplier α. The weight vector for predicting decision is $\beta = \sum_i \alpha_i x_i y_i; 0 \leq \alpha_i \leq C$. The instances x_i *with* $\alpha_i > 0$ are called support vectors, as they uniquely define the maximum margin hyperplane.

VII. Simulation Results

The experiments are conducted using UCF50 action dataset [22]. UCF50 is an activity recognition data set with 50

activities classes, composing of real Youtube videos. The large variations in cluttered background, camera motion, object scale, object appearance and pose, illumination conditions and viewpoint make the dataset to be very challenging. The total videos in UCF50 are 6680. The videos in UCF50 are grouped into 25 groups. For each group, the video clips have similar features, such as the same person, similar viewpoint, similar background, and so on. The classes or activities are visually shown in Fig. 6. The experiments are implemented on

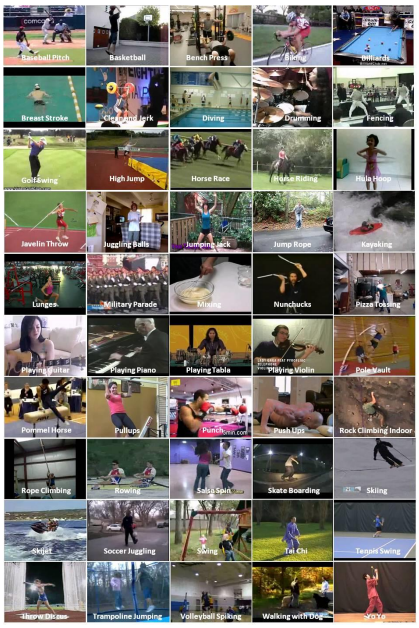


Fig. 6: UCF50 Dataset

computer with CPU i7, 2.6 GHz, 16 RAM, Matlab 2013b and R-Studio. Initially speaking, The features in UCF50 dataset are extracted using the spatiotemporal orientation energy, then the extracted values are described in vectors using template matching as action bank. The length of feature vector is 14746 and the number of observations is 6680. The R-Squared model is implemented to select the features that describe the variations in target. The features that explain the target class are selected and the other features are redundant or irrelevant. The minimum R-squared in our implementation is 0.005. It specifies the lower bound for the individual R-square value of a feature in order to be eligible for the model selection process. The number of selected features for each action is described in Fig. 7. The average number of features using R-Squared is 99 which is 0.67% from the original data. About 99.33% of features can't improve the performance of the model, but these features degrade negatively the recognition due to the large number of features which are redundant or irrelevant. The irrelevant features can make an over-fitting in the model.

The UCF50 features data are evaluated using 5-fold group-wise cross-validation, 5-fold video-wise cross-validation and $\frac{1}{3}(34\%)$ testing data. In our model, One-vs-rest SVM is applied to classify the actions using Linear kernel. The penalty is 1 and the maximum iterations is 25. For each action, positive video clips are labeled as 1 and negative videos are as labeled -1. For each action, R-Squared and SVM are applied. The accuracies are sorted for each action using 5-fold group-wise

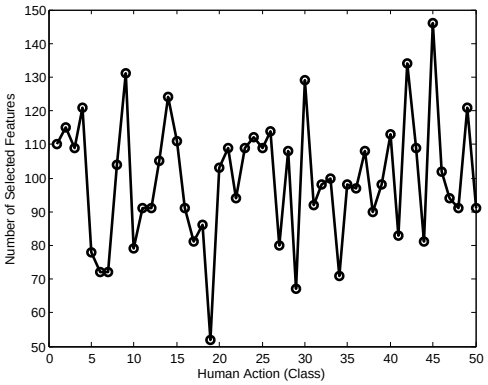


Fig. 7: Number of Selected Features using R-Squared Feature Selection for each Action

TABLE I: Sorted Accuracies score per class(action) in UCF50 dataset

2/3 train, 1/3 test		5-fold video-wise cross validation		5-fold group-wise cross validation	
Billiards	0.99	JumpingJack	0.9887	BreastStroke	0.956
PushUps	0.987	Billiards	0.98	Billiards	0.95
JumpRope	0.971	BreastStroke	0.965	CleanAndJerk	0.923
Biking	0.97	PlayingViolin	0.965	HighJump	0.891
BreastStroke	0.97	CleanAndJerk	0.964	JumpingJack	0.89
BaseballPitch	0.967	JugglingBalls	0.9549	PommelHorse	0.887
JumpingJack	0.965	PushUps	0.9528	PushUps	0.885
Mixing	0.96	Mixing	0.95	PlayingGuitar	0.881
PlayingViolin	0.958	PommelHorse	0.943	ThrowDiscus	0.877
YoYo	0.945	BaseballPitch	0.94	GolfSwing	0.875
Swing	0.938	MilitaryParade	0.937	PlayingPiano	0.873
Drumming	0.937	SalsaSpin	0.936	Mixing	0.872
RockClimbingIndoor	0.936	HighJump	0.9349	JumpRope	0.871
Fencing	0.935	PullUps	0.9333	MilitaryParade	0.864
PlayingPiano	0.933	Rowing	0.93	HorseRiding	0.863
HulaHoop	0.9303	Kayaking	0.9299	TaiChi	0.86
PommelHorse	0.93	GolfSwing	0.9295	BaseballPitch	0.857
PullUps	0.928	Nunchucks	0.9233	Fencing	0.854
HorseRace	0.925	RopeClimbing	0.9192	HorseRace	0.853
VolleyballSpiking	0.925	PlayingPiano	0.919	SkateBoarding	0.846
MilitaryParade	0.921	JumpRope	0.9189	BenchPress	0.845
Diving	0.92	RockClimbingIndoor	0.9189	PlayingViolin	0.845
Lunges	0.919	PlayingGuitar	0.9156	Skijet	0.84
HighJump	0.918	Diving	0.915	VolleyballSpiking	0.836
Kayaking	0.917	JavelinThrow	0.9102	PoleVault	0.835
Rowing	0.917	HorseRace	0.9094	Diving	0.831
JugglingBalls	0.916	Fencing	0.909	Punch	0.831
BenchPress	0.91	Biking	0.9069	RockClimbingIndoor	0.83
Skiing	0.91	Punch	0.906	SalsaSpin	0.827
Punch	0.909	HorseRiding	0.9056	Biking	0.823
HorseRiding	0.904	VolleyballSpiking	0.905	JugglingBalls	0.821
SalsaSpin	0.903	Swing	0.894	YoYo	0.818
ThrowDiscus	0.902	Drumming	0.891	JavelinThrow	0.813
CleanAndJerk	0.9	Lunges	0.89	Swing	0.807
RopeClimbing	0.9	Skijet	0.89	Basketball	0.789
GolfSwing	0.898	ThrowDiscus	0.8816	Drumming	0.781
JavelinThrow	0.891	SkateBoarding	0.879	PlayingTabla	0.781
Nunchucks	0.883	BenchPress	0.875	WalkingWithDog	0.778
SkateBoarding	0.88	TaiChi	0.875	Rowing	0.774
TrampolineJumping	0.877	Basketball	0.8723	PullUps	0.765
TaiChi	0.872	PoleVault	0.8687	Lunges	0.763
PlayingGuitar	0.863	PlayingTabla	0.8669	SoccerJuggling	0.756
PlayingTabla	0.861	Skiing	0.864	Nunchucks	0.753
Basketball	0.855	YoYo	0.859	RopeClimbing	0.753
SoccerJuggling	0.851	HulaHoop	0.84	HulaHoop	0.742
Skijet	0.843	PizzaTossing	0.8377	TennisSwing	0.739
PizzaTossing	0.836	SoccerJuggling	0.83	Kayaking	0.736
TennisSwing	0.802	TennisSwing	0.826	PizzaTossing	0.731
PoleVault	0.799	WalkingWithDog	0.796	Skiing	0.731
WalkingWithDog	0.742	TrampolineJumping	0.794	TrampolineJumping	0.721

cross-validation, 5-fold video-wise cross-validation and 1/3 testing data in Table I. The accuracies are visually shown in Fig. 8.

The overall accuracy using our approach is 82.64% for 5-fold group-wise cross-validation, 90.49% for 5-fold video-wise cross-validation and 90.8% for 34% testing data. The comparisons to available related works are described in Table II.

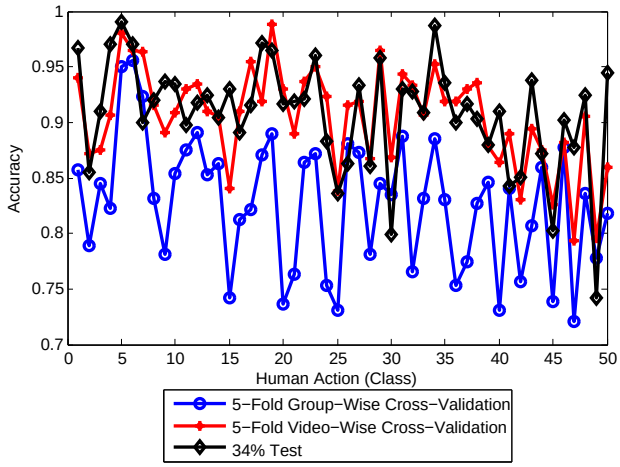


Fig. 8: Accuracy using R-Squared Feature Selection for each Human Action

TABLE II: Comparison with the Literature Results on UCF50 Dataset

Author	Experimental Setup	Accuracy
Our Method	5-fold group-wise cross validation	82.64%
Our Method	5-fold video-wise cross validation	90.49%
Our Method	2/3 training and 1/3 testing for each class	90.8%
Reddy and Shah [22]	Leave One Group Out Cross validation (25 cross-validations)	76.9%
Sadanand and Corso [14]	video-wise cross validation	76.4%
Sadanand and Corso [14]	group-wise cross validation	57.90%
Todorovic [23]	2/3 training and 1/3 testing for each class	81.03%
Solmaz et al. [24]	Leave One Group Out Cross validation(25 cross-validations)	73.70%
Kliper-Gross et al. [25]	Leave One Group Out Cross validation (25 cross-validations)	72.60%

VIII. Conclusions

Human activity recognition based on spatiotemporal orientation energy and activity template is simple and advanced discrimination techniques in detection and extraction features based on multiple activity detectors. The features in human activity recognition often more than the number of observations, so the feature selection is a major step before classification to avoid irrelevant or redundant features and over-fitting problems. R-Squared model is applied to get the best important discriminative features that explain the target. Also, R-Squared can handle a huge data in rapidly simplified manner. The model can significantly improve the performance/accuracy of human activities and reduce the features.

In the future, We will plan to apply the regression-based feature selection in human activity recognition based on different feature extraction methods that have large amount of features.

References

[1] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *ACM Computing Surveys (CSUR)*, vol. 43, April 2011.

[2] R. Gao, "Dynamic feature description in human action recognition," Master's thesis, Leiden Institute of Advanced Computer Science, Leiden University, 2009.

[3] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras, "Two-person interaction detection using body-pose features and multiple instance learning," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2012.

[4] A. G. Janecek, W. N. Gansterer, M. A. Demel, and G. F. Ecker, "On the relationship between feature selection and classification accuracy," in *JMLR: Workshop and Conference Proceedings*, pp. 90–105, 2008.

[5] E. Tuv, A. Borisov, G. Runger, K. Torkkola, I. Guyon, and A. R. Saffari, "Feature selection with ensembles, artificial variables, and redundancy elimination," *JMLR*, 2009.

[6] P. Doll'ar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proceedings of the 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS 05)*, pp. 65 – 72, October 2005.

[7] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space time shapes," in *IEEE International Conference on Computer Vision (ICCV)*, pp. 1395 – 1402, 2005.

[8] A. Klaser, M. Marszaek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *British Machine Vision Conference (BMVC)*, pp. 995 – 1004, 2008.

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

[10] X. Wang, L. Wang, and Y. Qiao, "A comparative study of encoding, pooling and normalization methods for action recognition," in *Computer Vision ACCV 2012* (K. Lee, Y. Matsushita, J. Rehg, and Z. Hu, eds.), vol. 7726 of *Lecture Notes in Computer Science*, pp. 572–585, Springer Berlin Heidelberg, 2013.

[11] K. Cheung, S. Baker, and T. Kanade, "Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. I–77–I–84 vol.1, June 2003.

[12] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence,*, vol. 29, pp. 2247 – 2253, Dec 2007.

[13] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates,," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257–267, 2001.

[14] S. Sadanand and J. Corso, "Action bank: A high-level representation of activity in video," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1234–1241, June 2012.

[15] L. Li, H. Su, Y. Lim, and F. Li, "Object bank: An object-level image representation for high-level visual recognition," *International Journal of Computer Vision*, vol. 107, no. 1, pp. 20–39, 2014.

[16] K. Derpanis, M. Sizintsev, K. Cannons, and R. Wildes, "Efficient action spotting based on a spacetime oriented structure representation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1990–1997, June 2010.

[17] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 2169–2178, 2006.

[18] M. Ramaswami and R. Bhaskaran, "A study on feature selection techniques in educational data mining," *Journal of Computing*, vol. 1, 2009.

[19] S. Garca, J. Luengo, and F. Herrera, *Data Preprocessing in Data Mining*. Springer International Publishing, 2015.

[20] H. Mazaar, E. Emary, and H. Onsi, "Evaluation of feature selection on human activity recognition," in *IEEE Seventh International Conference on Intelligent Computing and Information Systems (ICICIS15)*, pp. 105–113, 2015.

[21] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. Springer-Verlag New York, 2013.

[22] K. Reddy and M. Shah, "Recognizing 50 human action categories of web videos," *Machine Vision and Applications*, vol. 24, no. 5, pp. 971–981, 2013.

[23] S. Todorovic, "Human activities as stochastic kronecker graphs," in *Computer Vision ECCV 2012* (A. Fitzgibbon, S. Lazebnik, P. Perona,

Y. Sato, and C. Schmid, eds.), vol. 7573 of *Lecture Notes in Computer Science*, pp. 130–143, Springer Berlin Heidelberg, 2012.

[24] B. Solmaz, S. Assari, and M. Shah, "Classifying web videos using a global video descriptor," *Machine Vision and Applications*, vol. 24, no. 7, pp. 1473–1485, 2013.

[25] O. Kliper-Gross, Y. Gurovich, T. Hassner, and L. Wolf, "Motion interchange patterns for action recognition in unconstrained videos," in *European Conference on Computer Vision (ECCV)*, Oct. 2012.