Dictionary based action video classification with action bank

Shyju Wilson, M. Srinivas and C. Krishna Mohan
Department of Computer Science and Engineering
Indian Institute of Technology Hyderabad
Andhra Pradhesh, India, 502205
cs10p006@iith.ac.in, cs10p002@iith.ac.in, ckm@iith.ac.in

***Abstract*—Classifying action videos became challenging problem in computer vision community. In this work, action videos are represented by dictionaries which are learned by online dictionary learning (ODL). Here, we have used two simple measures to classify action videos, reconstruction error and projection. Sparse approximation algorithm LASSO is used to reconstruct test video and reconstruction error is calculated for each of the dictionaries. To get another discriminative measure projection, the test vector is projected onto the atoms in the dictionary. Minimum reconstruction error and maximum projection give information regarding the action category of the test vector. With action bank as a feature vector, our best performance is 59.3% on UCF50 (benchmark is 57.9%), 97.7% on KTH (benchmark is 98.2%)and 23.63% on HMDB51 (benchmark is 26.9%).**

***Index Terms*—Dictionary learning; Reconstruction error; Action videos;**

I. Introduction

In the digital world, exponential growth of video data has been observed for many years. This enforces the necessity of video understanding for efficient indexing and retrieval. The knowledge about actions in the video can be used as a discriminative information to organize the videos. Action bank[1], a high level representation of video, consists of output of many action detector that each give correlation volume. The rich semantic information in the action bank helps to improve performance of classifier. In this work, we have used action bank of each video as feature. Classification based on sparse representation of atoms in the dictionary has been emerged as a popular research in recent years. In sparse approximation, each signal can be represented as linear combination of fewer number of basis vectors in the known dictionary. The optimal approximation of members in vector space by using linear combination of fewer number of vectors in the known dictionary has come under the literature of compressive sensing [2]. Sparse representation gives better representation of the input signal and it is widely applied in many important fields like action recognition, object tracking, video super resolution, gesture recognition, face recognition, face hallucination etc. Sparse coding problem can be written as

$$\min\|x\|_0 \text{ such that } Dx = y, \tag{1}$$

where D is dictionary of atoms or signals, x is a sparse vector, y is the signal to be approximated and $\|\|_0$ denotes l_0 norm which looks for minimum number of non zero elements. Normally dictionary D is underdetermined system of linear equation ie. fewer number of equations than unknowns. In (1), it looks for solution x which is having minimum number of non zero elements.

There are several algorithms for sparse solution which include method of frames (MOF), basis pursuit, matching pursuit etc. [3]. LASSO (Least Absolute Shrinkage and Selection Operator) [4] is the variation of basis pursuit. It is a sparse approximation problem rather than sparse representation like basis pursuit. We have used LASSO sparse approximation algorithm to generate sparse vector from the learned dictionary for the reconstruction of input signal. After the reconstruction of input signal, reconstruction error to be calculated for classifying action videos. Tanaya Guha and Rabab Ward [5] consider reconstruction error as discriminative measure for classification of image and video, but in our proposed approach, we have used not only reconstruction error but also considered projection of the test vector onto the dictionary atoms to improve classification performance. If the test vector is more related to particular dictionary, then the length of projection becomes maximum.

Dictionary learning helps to reduce the size of the dictionary and it minimizes overall computational cost of classification. The method of optimal directions (MOD) was proposed by Engan et al. [6]. This was the first attempt to learn the dictionary which is also called sparsification process. For the given input signals $Y = [y_1 y_2 \dots y_N]$, MOD updates dictionary $D = [d_1 d_2 \dots d_K] \in R^{n \times K}$ and sparse matrix $X = [x_1 x_2 \dots x_N]$ alternatively which minimizes the representation error in (2),

$$\underset{D,X}{\operatorname{argmin}} \|Y - DX\|_F^2 \quad \text{subject to} \quad \forall i \ \ \|x_i\|_0 \leq T. \tag{2}$$

For learning the dictionary, MOD alternates sparse coding and dictionary update steps. This optimization problem is highly non-convex and may end up in local minimum. This is efficient method, but the problem with MOD is to find pseudo inverse while dictionary updating which leads to high computation. The K-SVD [7] is another dictionary learning algorithm similar to MOD except dictionary updation which includes singular value decomposition. Online dictionary learning (ODL) [8] is the recent algorithm which can handle large training sets espe-

cially videos. Zahra et al [9] suggested optimum dictionary by sparsifying each training signal in the dictionary rather than sparsifying whole dictionary and then solving the optimization problem using the idea of smoothed l_0 norm. In this work, we used ODL to learn the dictionary of video data which will be discussed in section III-A.

II. Features

Low level and mid level features are widely used in action recognition. Semantically rich features became more relevant now a days for the efficient representation of videos. Action bank, a high level representation of videos, consists of output of many action detectors that give a correlation volume. In this work, we have used action bank features which have been used by Sadanand and Corso in their work [1]. For better motion representation to detect unusual events, Wang and Liu [10] suggested random local feature (RLF) to the describe the spatio-temporal information of depth image. Jargalsaikhan [11] constructs 3D volume along sparse motion trajectories instead of dense trajectories and extract different features like histogram of oriented gradient (HOG), histogram of optical flow (HOF), motion boundary histogram (MBH), trajectory descriptor (TD) etc. to create bag of features (BoF). Wang et al. proposed high level concept action unit. The context aware descriptor that incorporates information from neighbouring interest points. Action unit is derived from the context aware descriptor using graph regularized non negative matrix factorization.

III. Proposed approach

We propose dictionary based approach for the classification of action videos with action bank. Each action video category is represented as a dictionary. These dictionaries are learned before classification. Dictionary learning helps to reduce the size of the input dictionary and provides better representation for each of the action categories. Sparse approximation plays an important role in every dictionary learning algorithms. In sparse approximation[12], the actual signal is to be approximated from the elementary signals in the dictionary called atoms, i.e. reconstruction of members in the vector space by using linear combination of fewer number of vectors in the dictionary. In the equation $Dx = y$, where D is a matrix with m number of rows and n number of columns. The given signal $y \in R^m$ is to be approximated by smallest possible number of vectors in the dictionary D such that $x \in R^n$ has least number of non-zero components. This is called sparse coding which is used in both dictionary learning and reconstruction of the input signal.

A. Online dictionary learning

Learning the dictionary will reduce redundancy, size of the dictionary, and improve the computational speed. We have set of input signal $Y = \{y_i\}_{i=1}^N$ which will be taken for training dictionary of size K ($K \ll N$). Here, these input signals are nothing but feature vectors. Online dictionary learning (ODL) [8] is computationally very effective and able to handle large datasets primarily in the field of image and video processing. Similar to other dictionary learning algorithms, ODL alternates two steps: sparse coding and dictionary update. For the sparse coding stage, the sparse vector is given by (3).

$$\underset{x \in R^K}{\operatorname{argmin}} \|y - Dx\|_F^2 + \lambda \|x\|_1, \tag{3}$$

where $D \in R^{m \times K}$ is initialized at the begining, $x \in R^K$ is the sparse vector, and λ is a regularization parameter. The input vector $y \in R^m$ and dictionary D are used to find sparse vector x. The new sparse vector x will be applied in (4) to get new dictionary. Each atom in the dictionary is updated using block coordinate descent method [8] and new dictionary is chosen by minimizing (4).

$$\underset{D \in C}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \|y_i - Dx_i\|_F^2 + \lambda \|x_i\|_1, \tag{4}$$

where C is the convex set of matrices with the following constraint:

$$C \doteq \{D \in R^{m \times K} \text{ s.t. } \quad \forall_i = 1, \ldots, k \quad d_i^T d_i \leq 1\}.$$

B. Action video classification

Videos are basically time series data. There are many classical approaches for classification of time series data such as hidden markov model (HMM) [13], dynamic time wrapping (DTW) [14], move split merge (MSM) [15]. In [16], Zhang et al work with human action recognition using sparse coding spatial pyramid matching. Spatio temporal interest points (STIP) from video sequence are projected onto three orthogonal planes to preserve the layout of STIPs. In this work, each action video is represented as action bank. Sparse coding is used to reconstruct the test vector and calculate reconstruction error for the classification. Reconstrcution error of sparse representation has been used to detect unusual events in surveillance applications [10]. K-SVD dictionary learning algorithm is used to learn the dictionaries and orthogonal matching pursuit (OMP) is used to get the sparse vector. Whenever reconstruction error exceeds predicted threshold, which is determined from the training data, then anomaly is detected.

The learned dictionaries are used for classification of action videos. Two diffrent measures, reconstruction error and projection, are applied to get discriminative information. The dictionary D is the concatenation of all learned dictionaries of each action video category. Consider there are m action categories, then the dictionary D becomes:

$$D = [d_{1,1} \ldots d_{1,n}, d_{2,1} \ldots d_{2,n} \ldots \ldots d_{m,1} \ldots d_{m,n}]$$

$D_k = \{d_{k,1} \ldots d_{k,n}\}$ denotes learned dictionary of k^{th} action category which is learned to n column vectors or dictionary atoms. The test vector y is approximated as linear combination atoms in the dictionary D_k ie. $y \approx D_k x_k$, x_k is the sparse vector. Here, the sparse approximation algorithm LASSO

generates sparse vector x_k from each of the dictionaries for the test vector y.

$$x_k = \text{LASSO}(D_k, y), \qquad k = 1 \dots m \tag{5}$$

Using the sparse vector x_k, the input vector y is reconstructed as the linear combination atoms in the dictionary.

$$y \approx D_k x_k, \qquad k = 1 \dots m \tag{6}$$

$D_k x_k$ is the reconstructed vector of test vector y from the dictionary D_k. Then the reconstruction error r_k becomes:

$$r_k = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \|y_i - D x_i\|_F^2 + \lambda \|x_i\|_1, \qquad k = 1 \dots m \tag{7}$$

$\text{R} = [r_1, r_2, \dots, r_m]^T$ contains reconstruction errors of y from m dictionaries. The minimum reconstruction error indicates action category of the test vector y.

Projection is another measure used here for classification. The test vector y is projected on to each of the dictionaries for the classification of action videos. The projection matrix P_i of each dictionary is constructed as follows:

$$P_i = D_i (D_i^T D_i)^{-1} D_i^T. \tag{8}$$

This projection matrix P_i is used to project test vector y onto the dictionary D_i. Then norm of the projection of test vector y can be considered as discriminative measure for the classification. p_i is norm of projection of y onto dictionary D_i:

$$p_i = \|P_i \, y\|_2, \tag{9}$$

$\text{P} = [p_1, p_2, \dots, p_m]^T$ contains norm of projection of y onto m dictionaries. Maximum projection indicates more correlation of test vector y to the vector space generated by the dictionary atoms of the corresponding dictionary.

Both reconstruction error and projection are together also used for classification by assigning weightage to them. For this purpose, the reconstruction vector R to be sorted in ascending order and projection vector P to be sorted in descending order. Lowest reconstruction error and highest projection are awarded maximum weightage. Final score is calculated by adding corresponding weights for classification decision making. Suppose we have 5 action categories, action A, action B, action C, action D and action E. Corresponding reconstruction error vector $\text{R} = [r_A, r_B, r_C, r_D, r_E]^T$ and projection vector $\text{P} = [p_A, p_B, p_C, p_D, p_E]^T$. After sorting R in ascending order and P in descending order, weightage is assigned to both R and P as shown below:

R	weightage	P	weightage
r_B	5	p_D	5
r_C	4	p_C	4
r_E	3	p_E	3
r_D	2	p_B	2
r_A	1	p_A	1

Final score of each class is calculated by adding corresponding weights.

Category	Final score
A	2
B	7
C	8
D	7
E	6

The action category belong to the maximum score will be assigned to test vector y. In the above example, action C is assigned to test vector y. This approach tried to reduce error occur in reconstruction error and projection. The intuition is that, the actual action category of test vector will always reside among top of the sorted vectors of R and P.

IV. Experimental study

The experiments are conducted with standard datasets KTH, UCF50 and HMDB51. In our experiment, action videos are classified in 3 ways viz.reconstruction error based, projection based and weightage method. In reconstruction error, action category belonging to minimum reconstruction error is assigned to test video. In projection, action category belonging to maximum projection is assigned to test video. In the third method, total score is calculated as explained in section III-B and then action category belonging to maximum score is assigned to test video. The KTH, UCF50, HMDB51 dataset have 6, 50, 51 action videos respectively. In all datasets, 2/3$^{\text{rd}}$ action videos of each category have been used for creating dictionary which is learned by online dictionary learning (ODL). Action bank is used as feature vector for the dictionary.

TABLE I
Confusion matrix of performance in KTH dataset

	boxing	clapping	handwaving	jogging	running	walking
boxing	1	0	0	0	0	0
clapping	0	0.94	0.06	0	0	0
handwaving	0	0.08	0.92	0	0	0
jogging	0	0	0	1	0	0
running	0	0	0	0	1	0
walking	0	0	0	0	0	1

UCF50 and HMDB51 are more challenging and realistic dataset compared to KTH. For UCF50, best performance achieved is 59.3% which is better than existing work's performance, 57.9%, using action bank [1]. This shows that dictionary is able to represent action videos efficiently. The performance of other datasets are also reasonably good, 97.7% on KTH (benchmark is 98.2%) and 23.6% on HMDB51 (benchmark is 26.9%). Detailed classification result of KTH, UCF50, HMDB51 are shown in table I, table II and table III, respectively. Overall performance of each experiment in different dataset shown in table IV.

V. Conclusions

The more challenging datasets make difficult action video classification. In this work, we have proposed dictionary

TABLE II
PERFORMANCE OF EACH ACTION CATEGORY OF UCF50 IN SORTED ORDER

Punch	0.96	HulaHoop	0.68	JugglingBalls	0.47
Billiards	0.94	Drumming	0.68	Swing	0.47
JumpingJack	0.93	Fencing	0.68	BaseballPitch	0.46
BenchPress	0.89	Kayaking	0.67	TennisSwing	0.45
HorseRiding	0.88	PullUps	0.63	VolleyballSpiking	0.45
HorseRace	0.86	Basketball	0.62	PlayingViolin	0.42
ThrowDiscus	0.84	Nunchucks	0.62	PizzaTossing	0.42
Mixing	0.83	HighJump	0.61	Biking	0.42
JumpRope	0.80	PushUps	0.57	SalsaSpin	0.41
RockClimbingIndoor	0.80	PlayingTabla	0.56	Diving	0.41
SkateBoarding	0.78	TaiChi	0.55	RopeClimbing	0.35
PlayingGuitar	0.77	MilitaryParade	0.52	PoleVault	0.28
PommelHorse	0.76	JavelinThrow	0.51	WalkingWithDog	0.27
BreastStroke	0.73	SoccerJuggling	0.50	TrampolineJumping	0.26
CleanAndJerk	0.70	YoYo	0.50	Lunges	0.26
GolfSwing	0.70	Rowing	0.49	Skijet	0.15
PlayingPiano	0.69	Skiing	0.48		

TABLE III
PERFORMANCE OF EACH ACTION CATEGORY OF HMDB51 IN SORTED ORDER

catch	0.71	ride_bike	0.32	kick_ball	0.21	eat	0.08
golf	0.60	push	0.32	hug	0.21	climb_stairs	0.08
laugh	0.60	turn	0.31	run	0.18	dive	0.07
walk	0.56	climb	0.31	cartwheel	0.17	sword_exercise	0.07
smile	0.50	talk	0.30	flic_flac	0.17	wave	0.06
pour	0.46	draw_sword	0.29	sit	0.17	shoot_gun	0.03
ride_horse	0.45	hit	0.29	dribble	0.15	somersault	0.02
pullup	0.41	jump	0.28	sword	0.14	kick	0.00
brush_hair	0.40	kiss	0.26	stand	0.14	punch	0.00
situp	0.40	shake_hands	0.26	smoke	0.11	shoot_ball	0.00
pushup	0.35	fencing	0.24	fall_floor	0.09	swing_baseball	0.00
clap	0.35	drink	0.22	pick	0.09	throw	0.00
shoot_bow	0.32	handstand	0.22	chew	0.08		

TABLE IV
OVERALL CLASSIFICATION PERFORMANCE (FIGURES IN %)

Classifier	KTH	UCF50	HMDB51
SVM[1]	98.20	57.90	26.90
Reconstruction Error	97.22	55.74	22.64
Projection	**97.69**	**59.30**	18.60
Weighted method	97.22	56.49	**23.62**

based classification of action videos with two discriminative measures, reconstruction error and projection. The dictionaries are learned by ODL for effective classification. Dictionary learning helps to reduce the size and redundancy of the dictionary. Action bank, high level feature, used here to represent videos. We have conducted three approaches for the classification of action videos, viz. reconstruction error based, projection based and weighted method. Our experiment shows learned dictionaries can effectively represent action videos and computationally effective. We can expect more performance from learned dictionaries by improving the classification methods. This will be considered in future work.

REFERENCES

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

[2] D. Donoho, "Compressed sensing," *Information Theory, IEEE Transactions on*, vol. 52, no. 4, pp. 1289 –1306, april 2006.

[3] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1998. [Online]. Available: http://epubs.siam.org/doi/abs/10.1137/S1064827596304010

[4] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288, 1994.

[5] T. Guha and R. Ward, "A sparse reconstruction based algorithm for image and video classification," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012, pp. 3601–3604.

[6] K. Engan, S. Aase, and J. Hakon Husoy, "Method of optimal directions for frame design," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, vol. 5, 1999, pp. 2443–2446 vol.5.

[7] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *Signal Processing, IEEE Transactions on*, vol. 54, no. 11, pp. 4311–4322, 2006.

[8] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ser. ICML '09. New York, NY, USA: ACM, 2009, pp. 689–696. [Online]. Available: http://doi.acm.org/10.1145/1553374.1553463

[9] Z. Sadeghipoor, M. Babaie-Zadeh, and C. Jutten, "Dictionary learning for sparse decomposition: A new criterion and algorithm," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, May 2013, pp. 5855–5859.

[10] C. Wang and H. Liu, "Unusual events detection based on multi-dictionary sparse representation using kinect," in *Image Processing (ICIP), 2013 20th IEEE International Conference on*, Sept 2013, pp. 2968–2972.

[11] I. Jargalsaikhan, S. Little, C. Direkoglu, and N. O'Connor, "Action recognition based on sparse motion trajectories," in *Image Processing (ICIP), 2013 20th IEEE International Conference on*, Sept 2013, pp. 3982–3985.

[12] D. Donoho, "Compressed sensing," *Information Theory, IEEE Transactions on*, vol. 52, no. 4, pp. 1289 –1306, april 2006.

[13] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, Feb 1989.

[14] J. Kruskall and M. Liberman, "The symmetric time warping algorithm: From continuous to discrete," *Time Warps*, pp. 125–162, 1983.

[15] A. Stefan, V. Athitsos, and G. Das, "The move-split-merge metric for time series," *IEEE Trans. on Knowl. and Data Eng.*, vol. 25, no. 6, pp. 1425–1438, Jun. 2013. [Online]. Available: http://dx.doi.org/10.1109/TKDE.2012.88

[16] L. Zhang, T. Wang, and X. Zhen, "Recognizing actions via sparse coding on structure projection," in *Image Processing (ICIP), 2013 20th IEEE International Conference on*, Sept 2013, pp. 2412–2415.