OBJECT MOVEMENT IDENTIFICATION VIA SPARSE REPRESENTATION

Haju Mohamed Mohamed Naleer

Department of Mathematics, Faculty of Applied Science
South Eastern University of Sri Lanka, Sammanthuari, Sri Lanka
hmmnaleer@gmail.com

ABSTRACT: Object Movement Identification from videos is very challenging, and has got numerous applications in sports evaluation, video surveillance, elder/child care, etc. In thisresearch, a model using sparse representation is presented for the human activity detection from the video data. This is done using a linear combination of atoms from a dictionary and a sparse coefficient matrix. The dictionary is created using a Spatio Temporal Interest Points (STIP) algorithm. The Spatio temporal features are extracted for the training video data as well as the testing video data. The K-Singular Value Decomposition (KSVD)algorithm is used for learning dictionaries for the trainingvideo dataset. Finally, human action is classified using aminimum threshold residual value of the corresponding actionclass in the testing video dataset. Experiments are conducted onthe KTH dataset which contains a number of actions. Thecurrent approach performed well in classifying activities with asuccess rate of 90%.

Keywords: sparse representation, human activity detection, KSVD, STIP,dictionary learning

1. INTRODUCTION

People's behavior is analyzed using various methods for human activity detection from videos. The video data forthis analysis is collected using wearable sensors, RGBDsensors or recorded videos. It is very challenging to findfeatures of human activity from video due to following reasons:

- Human activities are diverse as it can be performedby person of different size and appearance
- Different human activities share common movements
- Variability in the video data

Numerous researches have been done to develop different techniques in the area of human activity detection. One of the common techniques for human activity detection is the Hidden Markov Model (HMM) [1]. The HMM technique was coupled with other techniques to obtain improved methods such as Maximum Entropy Markov Model (MEMM). Li et al. developed algorithms based on this MEMM [2].

Sung et al. worked with MEMM and developed new algorithm to work with RGBD (Red Green Blue Depth) image [3]. Another approach is pervasive computing. Wilde collected data using pervasive sensors and used existing classification techniques to detect human activity in her thesis [4]. Recently, a number of feature based methods for action detection from videos are proposed in [5]. In feature based methods there are three main steps. The first step is to find the interest points. The second step is the feature acquisition. In the final step, the classification of actions is done using the features extracted from the video data. In spite of the success of different methods, sparse representation is getting lot of attention in computer vision and signal processing area. Zhang et al. proposed a sparse representation based human activity detection [6]. Authors collected data using wearable sensors, and then they extracted features from the data to form local feature vectors. This research effort will explore human activity detection using sparse representation from video data recorded in a controlled environment. Section II provides the background research, Section III discusses about the methodology, Section IV presents the simulation results, and Section V gives details about Conclusion and Future Work. This is followed by the References used for this research.

2. BACKGROUND REASERCH

Niu et al. presented an algorithm for detecting and recognizing human activities for outdoor surveillance applications [7]. This algorithm is built on top of low-level motion detection algorithms such as frame differencing and feature correlation. They used a representation of human activities based on tracked trajectories for activity recognition. For this purpose, the different interaction patterns among a group of people are distinguished. This is done by identifying the unique signatures of the relative position and from the velocity of the participants' trajectories. Saxena et al. has performed detection and recognition of human activity in the unstructured environments [3]. They used a Red Green Blue Depth (RGBD) sensor as the input sensor, and computed a set of features based on human pose and motion. Human activities have a natural hierarchical structure. The authors captured this hierarchical nature using a maximum entropy Markov model (MEMM). It is hard to capture variations in human activities using single graphical model. They presented a method of on-the fly graph structure selection that can automatically adapt to variations in the task speeds and style. Finally, they extracted features using the Prime Sense skeleton tracking system in combination with aspecially placed Histogram of Oriented Gradient (HOG) computer vision features. Yin et al. proposed an approach for abnormal activity detection based on sensor readings from wearable sensors [8]. It was hard to obtain a large amount of training data for abnormal activities, but it was possible for normal activities. This enabled the creation of well estimated models for normalactivities, which can be adapted for abnormal activities at a later stage. They proposed a two-phase approach for abnormal activity detection. In the first phase, they built a one-class Support Vector Machine (SVM) solely based on normal activities. This can filter out activities having a very high probability of being normal. Then further detection is done on the suspicious traces. In the second phase, they performed a kernel based nonlinear regression (KNLR) analysis to deriveabnormal activity models from a general normal activity model in an unsupervised manner.

3. METHODOLOGY ANALYSIS

A. Sparse representation

Sparse representation along with dictionary learning is used in many signal and image processing tasks such as image denoising, face recognition, image classification etc. The technique of finding a matrix with a small number of nonzero coefficients is called as Sparse Representation. It is possible to construct a model that is best suitable for the training data with a linear combination of a small number of elementary signals called atoms. These atoms are chosen from a dictionary, D. A dictionary (D) is a collection of atoms such that any signal can be represented by more than one combination of different atoms. Sparsity of a signal is measured using L-P norm for a given *p* that will give absolute value of every entry of '$\propto$' raised to 'p' power and add all of them together.

$$\|\alpha\|_p^p = \sum |\alpha_j|^p$$

Assuming dictionary (*D*) is fixed; a sparserepresentation of sample (*X*) is obtained by minimizing $\|\propto\|_0$in the linear equation in (1).

$$\alpha^* = argmin \|\alpha\|_0 \; s.t\; X = D\alpha \quad (1)$$

Where *'D'* is dictionary of size dictionary of size '*K*', '*X*' is a set of *N* input signals and '$\propto$' is the sparse matrix.

The sparse representation of signals is demonstrated in Fig 1.

$||\alpha||^0$ is the L_0 norm and it will give a number of nonzero components in vector *α*. The general problem of finding a representation with the smallest number of atoms from a dictionary has been shown to be Nondeterministic Polynomial-time hard (*NP*-hard). However, if certain conditions on sparsity are satisfied, i.e. if the solution is sparse enough, the sparse representation can be recovered by *L1*- minimization. This means that the equivalent solution can be obtained by replacing L_0norm in (1) with L_1 norm as shown in equation (2),

$$\alpha^* = argmin \|\propto\|_1 \; s.t\; X = D\alpha \qquad (2)$$

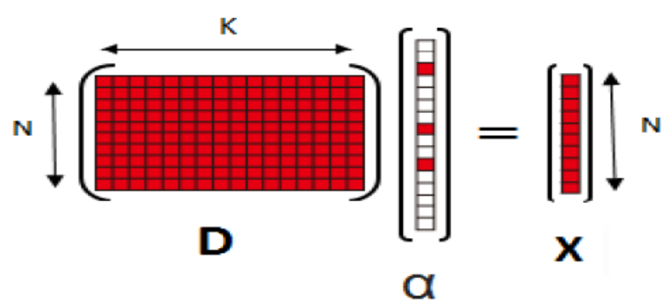

Figure 1. Sparse Representation

But sometimes the video data available for human action detection are noisy. It may not be possible to expresstest video data as sparse as training video data. So the equation(2) can be solved by a more stable and simple L1-minimizationproblem as given in equation (3),

$$\alpha^* = argmin \; \|\alpha\|_1 \; s.t \; X = \|D\alpha - X\| \leq \in \qquad (3)$$

where '€' is error tolerance.Equation (3) can be solved by using severalalgorithms. These algorithms are mainly classified into twodifferent types, namely *Greedy* algorithms and Relaxationalgorithms. Greedy algorithms iteratively build up the signalapproximation by taking one coefficient at a time; e.g.Matching pursuit, or Orthogonal Matching Pursuit. Inrelaxation method, algorithms process all coefficientssimultaneously; e.g. Basis pursuit, and *FOCal*Underdetermined System Solver (FOCUSS). Other than abovementioned methods, there are few more methods available tosolve L_1-minimization. One of them is *Homotopy* algorithms,and in this research the L_1 homotopy algorithm is used.

There are a number of steps that needs to becompleted for activity detection before solving the equation in(3). Initially, the features are extracted from the video. Thesefeatures are the input for dictionary learning. This dictionarywill be used as input for sparse coding. Finally, action oractivity classification is done using this dictionary. Thearchitecture diagram of Human Activity Detection usingSparse Representation (HADSR) system using this approach isshown in Fig 2.

Assuming that a set of videos in training set contains enough known actions, the aim is to learn activities from these
videos, and achieve classification of activities for the testingvideo data set.

B. Spatio Temporal Feature Extraction

The initial step is the extraction of human activity features from the training video data. For this purpose, the Spatio temporal features of the activities are used. There are anumber of methods available for extracting Spatio temporal features from the video such as SURF algorithm, local cuboidmethod using optical flow, low level motion features, Spatio Temporal Interest Points (STIP) etc. In this research, the Spatio Temporal Interest Points (STIP) method is used forboth the training set and the testing set.

Spatio Temporal Interest Points (STIP)

The Spatio Temporal Interest Points (STIP) algorithm detects the significant changes locally, both in spaceand time dimensions. The general idea of extracting spatiotemporal interest points from video is similar to extracting thespatial interest points. Instead of extracting features from animage, interest point detector should work on stack of images. The idea of detecting spatio temporal interest points are built upon the Harris and Forstner interest point operators[9] [10]. Laptev and Lindeberg extended this idea of interest points into spatio temporal domain, and illustrated how these resulting spatio temporal features often corresponds tointeresting events in video data [11]. This method detects Spatio Temporal Interest Points (STIP) and computes thecorresponding local space-time descriptors. The STIP detectspoints for a set of multiple combinations of spatial andtemporal scales. After detecting the interest points, thedescriptors can be detected using Histograms of OrientedGradients (HOG) or Histograms of Optical Flow (HOF)method. In this research the HOG method is used.

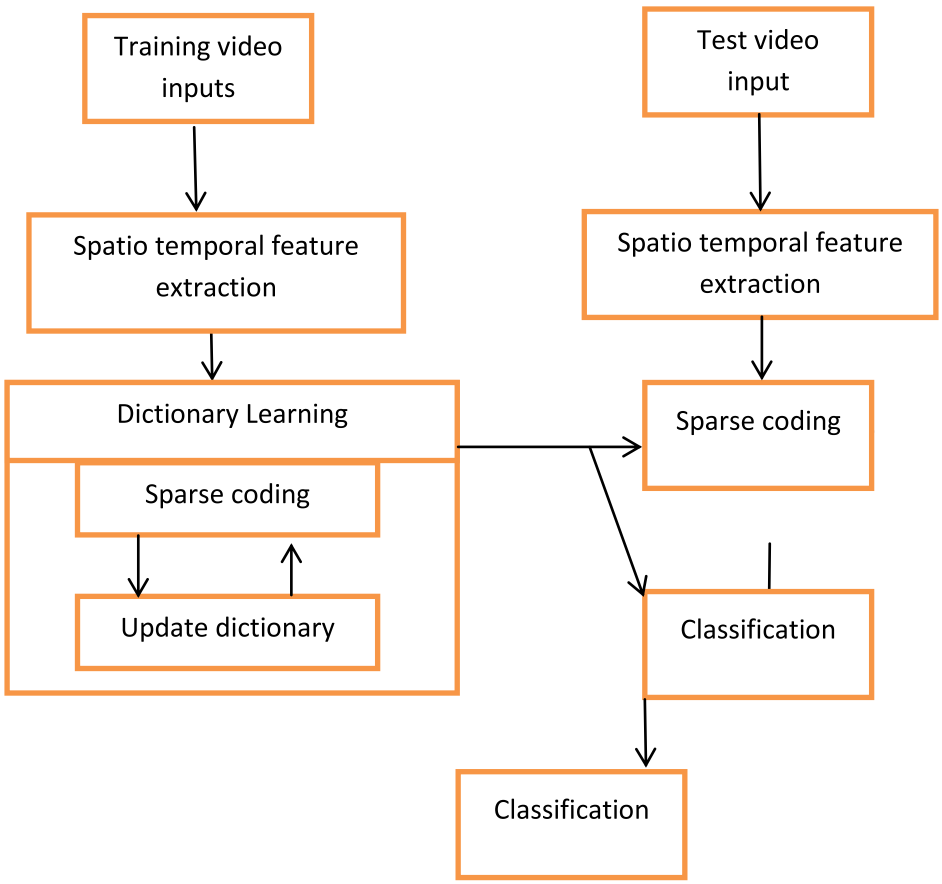

Figure 2. Architecture of HADSR system

C. Dictionary learning

The second step is the dictionary learning for everyaction class available in the training data. If there are jdifferent activity classes in training data, then create j numberof action specific dictionaries, D (sub-dictionaries). After reading action specific sub-dictionaries combine them all toform an over complete structured dictionary, D.jAn over complete dictionary D, that leads to sparserepresentations can either be pre-designed, or designed for a particular dataset by using its content. Choosing a predesigned dictionary is appealing because it is simpler and faster. But success of these dictionaries depends on how suitable they are for the test data. An over complete dictionary, D designed for a particular training data is more successful than a commonly used pre-designed dictionaries. This approach is used in this research.

K – Singular Value Decomposition (K-SVD) Algorithm

Recently, few researches have been done indictionary learning, mainly on the study of pursuit algorithms.In this research, the K – Singular Value Decomposition (KSVD)algorithm is used for learning an over completedictionary [12]. K-SVD

algorithm is a generalization of the Kmeansclustering process. This algorithm will create theDictionary (D), which will lead to the best possiblerepresentation of every member in the set with strict sparse
constraints. K-SVD is an iterative method that alternates betweensparse coding of the training data based on the currentdictionary and a process of updating the dictionary. The process for K-SVD algorithm is shown in Fig. 3.

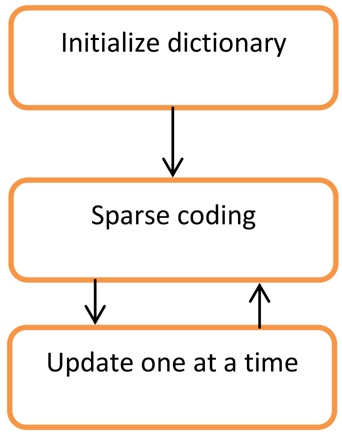

Figure 3. Processes in K-SVD algorithm

Initially, consider the sparse coding stage. Assume that dictionary, D is fixed, find a sparse representation with
the coefficients summarized in matrix X. Equation (3) can be rewritten as,

$$\|D\alpha - X\|_2^2 = \sum \|D\alpha_i - X_i\|_2^2 \;\; for\; i = 1,2,3, \dots \dots \dots \dots, N \qquad (4)$$

For updating the dictionary, assume X is fixed. Also,consider only one column in the dictionary, dand thecoefficients associated with it, which is the kth row in thesparse matrix, X. Equation (3) can be rewritten as,

$$\|D\alpha - X\|_2^2 = \sum \left\| d_j \alpha_k^j - X \right\|_2^2 \qquad (5)$$

K-SVD algorithm sweeps through all the columnsand will use the most recently updated values from theprevious step. Also, all updates in dictionary are done basedon the same X. In each sparse coding step, the totalrepresentation error decreases. During the dictionary updateprocess there will be changes in the representation error
without affecting the sparse constraints. The success of thisprocess is depending on whether the K-SVD algorithm is flexible enough to work with any pursuit algorithm such asOrthogonal Matching Pursuit (OMP), Basis pursuit (BP), orFOCal Underdetermined System Solver (FOCUSS).

D. Action Classification

The final step is the activity classification in the test videodata. Locations of non-zero coefficients of a* can be used toclassify these actions. Each non-zero coefficient of arepresenta correspondence of an action in training set to an action in thetesting video data set. Ideally, the non-zero coefficients shouldonly be associated with a training set which has the same classas the testing set. However non-zero coefficients are spreadacross more than one class. The action in the testing set can beidentified using the residual error (R). The residual error iscalculated using the equation (6).

$$R(q, \alpha^*) = \|q - D_i \alpha_i^*\| \qquad (6)$$

After calculating residual error (R) for every action, theaction in the test video (q) is classified to the action having
smallest residual error, R.

$$Lable(q) = argmin\, R(q, \alpha^*) \qquad (7)$$

4. SIMULATION RESULT

Experiments were conducted using the publiclyavailable KTH dataset. KTH is a commonly used dataset for action recognition. It contains 25 subjects performing 6actions in 4 different scenarios. The actions include walking, running, jogging, boxing, hand waving and hand clapping. After selecting a test video, the first step is the
extraction of spatio temporal features from the video. Thefeatures are found; the points are detected from a frame in thevideo as shown in Fig 4.

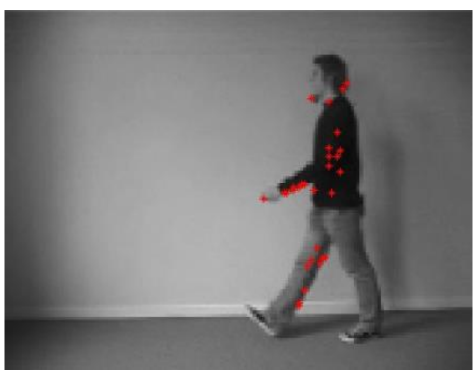

Figure 4. Detected spatio temporal features

Action specific dictionary is created using these features for each class. Using all the action specificdictionaries, an over complete dictionary is created. The over complete dictionary for walking is shown in Fig 5.

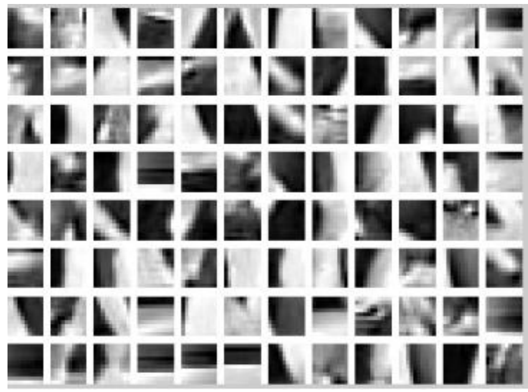

Figure 5. Dictionary learned using the KSVD algorithm

After creating the dictionary, equation (3) is solved using L_1- norm solver named L_1 homotopy. It provided a for every action class for the test video. The action classificationusing equation (6) is done to obtain the residual error of thetest video for every action. The residual errors obtained forrunning, walking and hand waving is shown in Table 1.

Table 1. Residual Error

Training/Test	Walking	Running	Hand waving
Walking	$1.23e^{+09}$	$1.50e^{+09}$	$1.46e^{+09}$
Running	$7.22e^{+09}$	$7.96e^{+07}$	$3.02e^{+08}$
Hand waving	$7.49e^{+08}$	$3.16e^{+08}$	$1.10e^{+08}$

Table I shows that the residual error for a corresponding class is minimal compared to any other residual errors. Label the action in the test video with the class ofminimum residual error. Simulation is done using this approach with 30 test videos of 3 different activitiesperformed by different subjects. Table II shows the result of the action classification.

Table II: Result of action classification

Test video	Walking	Running	Hand waving
Correctly classified	10	9	8
Misclassified	0	1	2

5. CONCLUSION AND FUTURE WORK

In this research, a sparse representation model forhuman activity detection is presented. Spatio Temporal Interest Points (STIP) is used for extracting spatio temporal features from the training video data as well as testing videodata. Action specific dictionaries are created using spatiotemporal features of training videos. This approach used KSVDalgorithm for learning dictionaries for a particulardataset. For solving sparse linear equation L-norm minimization is used. After solving the equation, the residualerrors are computed for the actions in the test video using anover complete dictionary. These residual errors are used toclassify the activity in the test video. Action classification isdone based on the minimal residual error to a class in thetraining set. The KTH dataset is used for the simulation, and ithas proven that this approach is successful in classifying theactivities very effectively with a success rate of 90%. Thisapproach works in a controlled environment and with less noisy (cluttered) videos. The research may be extended towork with any video with multiple persons and /or otherobjects present in the video.

6. REFERENCES

YAMATO, J., JUN OHYA, ISHII K., (1992) "Recognizing human action in time-sequential images using hidden Markov model," 1992 IEEEComputer Society Conference on Computer Vision and PatternRecognition (CVPR '92), Champaign, IL, USA, pp.379- 385.

SMINCHISESCU, C., KANAUJIA, A, ZHIGUO LI, METAXAS, D.,(2005) "Conditional models for contextual human motion recognition," Tenth IEEE International Conference on ComputerVision (ICCV 2005), vol.2, pp.1808-1815.

JAEYONG SUNG, PONCE C., SELMAN, B., SAXENA A., (2012) "Unstructured human activity detection from RGBD images," 2012 IEEE International Conference on Robotics and Automation (ICRA- 2012),St. Paul, Minnesota, USA, pp. 842- 849.

WILDE, ADRIANA GABRIELA.(2011) “Activity recognition for motion-awarepervasive systems”, Department of Informatics, Masters’ Thesis, University of Fribourg, Switzerland.

WANG H., ULLAH M., KLASER A., LAPTEV I. & SCHMID C.(2009) “Evaluation of local spatio-temporal features for action recognition”,Proc. British Machine Vision Conference (BMVC'09), London, UK.

MI ZHANG, WENYAOXU, SAWCHUK, A.A., SARRAFZADEH, M., (2012) "Sparse representation for motion primitive-based human activity modelingand recognition using wearable sensors," 21st InternationalConference on Pattern Recognition (ICPR) 2012, pp.1807-1810.

WEI NIU, JIAO LONG, DAN HAN, YUAN-FANG WANG,(2004) "Human activity detection and recognition for video surveillance", 2004 IEEEInternational Conference on Multimedia and Expo - 2004. ICME '04,vol.1, pp. 719-722.

JIE YIN, QIANG YANG, PAN, J.J.,(2008) "Sensor-Based Abnormal Human Activity Detection,” IEEE Transactions on Knowledge and Data Engineering, vol.20, no.8, pp.1082-1090.

W. FORSTNER & E. GULCH,(1987) “A fast operator for detection andprecise location of distinct points, corners and centres of circular features”, ISPRS Intercomission Conference on Fast Processing of Photogrammatic Data, Interlaken, pp. 281–305.

C. HARRIS & M. STEPHENS, (1988) “A combined corner and edge detector”, Proceedings of the Fourth Alvey Vision Conference, Manchester, UK, vol. 15, pp. 147–151.

I. LAPTEV & T. LINDEBERG,(2003) “Space-Time Interest Points, “Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV’03), Nice, France, pp. 432–439.

AHARON, M.; ELAD, M.; BRUCKSTEIN, A.,(2006) "K -SVD: An Algorithm for Designing Over complete Dictionaries for Sparse Representation," IEEE Transactions on Signal Processing, vol.54,no.11, pp.4311-4322.