OBJECT MOVEMENT IDENTIFICATION VIA SPARSE REPRESENTATION

Haju Mohamed Mohamed Naleer

Division of Computer Science, Department of Mathematics,
Faculty of Applied Science
South Eastern University of 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 this research, 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 training video dataset. Finally, human action is classified using a minimum threshold residual value of the corresponding action class in the testing video dataset. Experiments are conducted on the KTH dataset which contains a number of actions. The current approach performed well in classifying activities with a success rate of 90%.

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

INTRODUCTION

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

- Human activities are diverse as it can be performed by 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.

METHODOLOGY ANALYSIS

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 'α' raised to '*p*' power and add all of them together.

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

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

$$\alpha^* = argmin\|\alpha\|_0 \, s.t \, X = D\alpha \tag{1}$$

Where '*D*' is dictionary of size dictionary of size '*K*', '*X*' is a set of *N* input signals and 'α' 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_0 norm in (1) with L_1 norm as shown in equation (2),

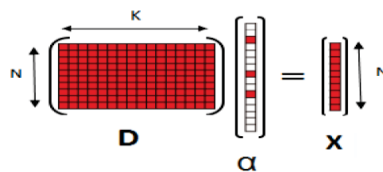$$\alpha^* = argmin\|\alpha\|_1 \, s.t \, X = D\alpha \tag{2}$$

Fig.1: Sparse Representation

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

$$\alpha^* = argmin\|\alpha\|_1 \, s.t \, X = \|D\alpha - X\| \leq \epsilon \tag{3}$$

where 'ϵ' is error tolerance. Equation (3) can be solved by using several algorithms. These algorithms are mainly classified into two different types, namely *Greedy*

algorithms and Relaxation algorithms. Greedy algorithms iteratively build up the signal approximation by taking one coefficient at a time; e.g. Matching pursuit, or Orthogonal Matching Pursuit. In relaxation method, algorithms process all coefficients simultaneously; e.g. Basis pursuit, and *FOCal* Underdetermined System Solver (FOCUSS). Other than above mentioned methods, there are few more methods available to solve 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 be completed for activity detection before solving the equation in (3). Initially, the features are extracted from the video. These features are the input for dictionary learning. This dictionary will be used as input for sparse coding. Finally, action or activity classification is done using this dictionary. The architecture diagram of Human Activity Detection using Sparse Representation (HADSR) system using this approach is shown 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 testing video data set.

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 a number of methods available for extracting Spatio temporal features from the video such as SURF algorithm, local cuboid method 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 for both 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 space and time dimensions. The general idea of extracting spatio temporal interest points from video is similar to extracting the spatial interest points. Instead of extracting features from an image, 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 to interesting events in video data [11]. This method detects Spatio Temporal Interest Points (STIP) and computes the corresponding local space-time descriptors. The STIP detects points for a set of multiple combinations of spatial and temporal scales. After detecting the interest points, the descriptors can be detected using Histograms of Oriented Gradients (HOG) or Histograms of Optical Flow (HOF) method. In this research the HOG method is used.
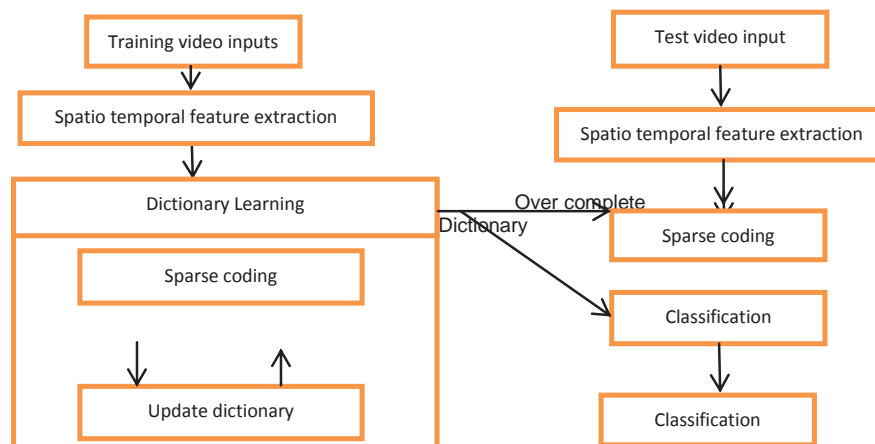

Fig. 2: Architecture of HADSR system

Dictionary learning

The second step is the dictionary learning for every action class available in the training data. If there are j different activity classes in training data, then create j number of action specific dictionaries, D (sub-dictionaries). After reating action specific sub-dictionaries combine them all to form an over complete structured dictionary, D. j An over complete dictionary D, that leads to sparse representations 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 in dictionary learning, mainly on the study of pursuit algorithms. In this research, the K – Singular Value Decomposition (KSVD) algorithm is used for learning an over complete dictionary [12]. K-SVD algorithm is a generalization of the Kmeans clustering process. This algorithm will create the Dictionary (D), which will lead to the best possible representation of every member in the set with strict sparse
constraints. K-SVD is an iterative method that alternates between sparse coding of the training data based on the current dictionary and a process of updating the dictionary. The process for K-SVD algorithm is shown in Fig 3.
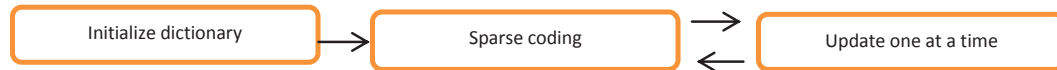

Fig. 3: Processes in K-SVD algorithm

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,

$$\|D\alpha - X\|_2^2 = \sum \|D\alpha_i - X_i\|_2^2 \ for \ i = 1,2,3,\ldots\ldots\ldots,N \tag{4}$$

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

$$\|D\alpha - X\|_2^2 = \sum \|d_j \alpha_k^j - X\|_2^2 \tag{5}$$

K-SVD algorithm sweeps through all the columns and will use the most recently updated values from the previous step. Also, all updates in dictionary are done based on the same X. In each sparse coding step, the total representation error decreases. During the dictionary update process there will be changes in the representation error
without affecting the sparse constraints. The success of this process is depending on whether the K-SVD algorithm is flexible enough to work with any pursuit algorithm such as Orthogonal Matching Pursuit (OMP), Basis pursuit (BP), or FOCal Underdetermined System Solver (FOCUSS).

D. Action Classification

The final step is the activity classification in the test video data. Locations of non-zero coefficients of a* can be used to classify these actions. Each non-zero coefficient of a represent a correspondence of an action in training set to an action in the testing video data set. Ideally, the non-zero coefficients should only be associated with a training set which has the same class as the testing set. However non-zero coefficients are spread across more than one class. The action in the testing set can be identified using the residual error (R). The residual error is calculated using the equation (6).

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

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

$$Lable(q) = argmin \, R(q, \alpha^*) \tag{7}$$

SIMULATION RESULT

Experiments were conducted using the publicly available KTH dataset. KTH is a commonly used dataset for action recognition. It contains 25 subjects performing 6 actions 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. The features are found; the points are detected from a frame in the video as shown in Fig 4. Action specific dictionary is created using these features for each class. Using all the action specific dictionaries, an over complete dictionary is created. The over complete dictionary for walking is shown in Fig 5. 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 classification using equation (6) is done to obtain the residual error of the test video for every action. The residual errors obtained for running, walking and hand waving is shown in Table 1.
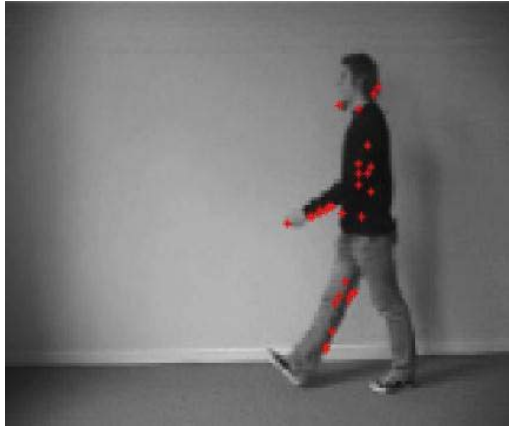

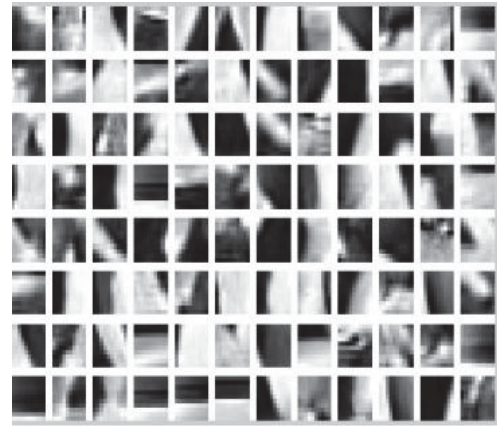Fig. 4: Detected spatio temporal features

Fig. 5: Dictionary learned using the KSVD algorithm

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 of minimum residual error. Simulation is done using this approach with 30 test videos of 3 different activities performed by different subjects. Table II shows the result of the action classification.

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 II: Result of action classification

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

CONCLUSION AND FUTURE WORK

In this research, a sparse representation model for human 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 video data. Action specific dictionaries are created using spatio temporal features of training videos. This approach used KSVD algorithm for learning dictionaries for a particular dataset. For solving sparse linear equation L-norm minimization is used. After solving the equation, the residual errors are computed for the actions in the test video using an over complete dictionary. These residual errors are used to classify the activity in the test video. Action classification is done based on the minimal residual error to a class in the training set. The KTH dataset is used for the simulation, and it has proven that this

approach is successful in classifying the activities very effectively with a success rate of 90%. This approach works in a controlled environment and with less noisy (cluttered) videos. The research may be extended to work with any video with multiple persons and /or other objects present in the video.

REFERENCES

[1.] Yamato, J.; Jun Ohya; Ishii K., "Recognizing human action in time-sequential images using hidden Markov model," 1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '92), Champaign, IL, USA, pp.379- 385, Jun 1518, 1992

[2.] Sminchisescu, C.; Kanaujia, A; Zhiguo Li; Metaxas, D., "Conditional models for contextual human motion recognition," Tenth IEEE International Conference on Computer Vision (ICCV 2005), vol.2, pp.1808-1815, Oct. 17-21, 2005

[3.] Jaeyong Sung; Ponce C.; Selman, B.; Saxena A., "Unstructured human activity detection from RGBD images," 2012 IEEE International Conference on Robotics and Automation (ICRA- 2012), St. Paul, Minnesota, USA, pp. 842- 849, May 14-18, 2012

[4.] Wilde, Adriana Gabriela. "Activity recognition for motion-aware pervasive systems", Department of Informatics, Masters' Thesis, University of Fribourg, Switzerland, 2011

[5.] Wang H.; Ullah M.; Klaser A.; Laptev I. and Schmid C. "Evaluation of local spatio-temporal features for action recognition",

[6.] Proc. British Machine Vision Conference (BMVC'09), London, UK, September 2009

[7.] Mi Zhang; Wenyao Xu; Sawchuk, A.A.; Sarrafzadeh, M., "Sparse representation for motion primitive-based human activity modeling and recognition using wearable sensors," 21st International Conference on Pattern Recognition (ICPR) 2012, pp.1807-1810, 1115 Nov. 2012

[8.] Wei Niu; Jiao Long; Dan Han; Yuan-Fang Wang, "Human activity detection and recognition for video surveillance", 2004 IEEE International Conference on Multimedia and Expo - 2004. ICME '04, vol.1, pp. 719-722, 30 June 2004

[9.] Jie Yin; Qiang Yang; Pan, J.J., "Sensor-Based Abnormal HumanActivity Detection," IEEE Transactions on Knowledge and Data Engineering, vol.20, no.8, pp.1082-1090, Aug.2008

[10.] W. Forstner and E. Gulch, "A fast operator for detection and precise location of distinct points, corners and centres of circular features", ISPRS Intercomission Conference on Fast Processing of Photogrammatic Data, Interlaken, pp. 281–305, 1987

[11.] C. Harris and M. Stephens, "A combined corner and edge detector", Proceedings of the Fourth Alvey Vision Conference, Manchester, UK, vol. 15, pp. 147–151, 1988.

[12.] Laptev and T. Lindeberg, "Space-Time Interest Points,"Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV'03), Nice, France, pp. 432–439, 2003

[13.] Aharon, M.; Elad, M.; Bruckstein, A., "K -SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation," IEEE Transactions on Signal Processing, vol.54,no.11, pp.4311-4322, Nov. 2006