Human Action Recognition System

Submitted By

Jonti Talukdar (14BEC057)

Bhavana Mehta (14BEC028)

ELECTRONICS AND COMMUNICATION

INSTITUTE OF TECHNOLOGY

NIRMA UNIVERSITY AHMEDABAD-382481

APRIL 2017

Human Action Recognition System

Idea Lab Project

Submitted By

Jonti Talukdar (14BEC057)

Bhavana Mehta (14BEC057)

Under the mentorship of

Dr. Tanish Zaveri

ELECTRONICS AND COMMUNICATION

INSTITUTE OF TECHNOLOGY

NIRMA UNIVERSITY AHMEDABAD-382481

APRIL 2017

Declaration

We do hereby declare that the technical project report submitted is original, and is the outcome of the independent investigations/research carried out by us and contains no plagiarism. The research is leading to the discovery of new facts/techniques/correlation of scientific facts already known. This work has not been submitted to or supported by any other University or funding agency.

We do hereby further declare that the text, diagrams or any other material taken from other sources (including but not limited to books, journals and web) have been acknowledged, referred and cited to the best of our knowledge and understanding.

Date:

Place: Ahmedabad

Signature of Student 1	Signature of Student 2	Signature of Mentor
Jonti Talukdar	Bhavana Mehta	Dr. Tanish Zaveri
14BEC057	14BEC028	

NIRMA UNIVERSITY INSTITUTE

OF TECHNOLOGY IDEA LAB

ELECTRONICS AND COMMUNICATION ENGINEERING

Annual/Final Report of the work done on the Idea Lab Project. (Report to be submitted within 3 weeks after completion of the project)

1. Idea Lab Project ID:

2. Project Title: Human Action Recognition System

3. Period of Project: April 2016 to April 2017

4. (a). Name of Student (Roll No.): Jonti Talukdar (14BEC057)
   Department: EC

   (b). Name of Student (Roll No.): Bhavana Mehta (14BEC028)
   Department: EC

   (c). Name of Mentor: Dr. Tanish Zaveri.

5. Project Start Date: May 2016.

6. (a) Total Amount Approved: Rs. 17,000/-

   (b) Total Expenditure: Rs. 14,049/-

   (c) Report of the work done:

   i. Brief objective of the project: To design a Real time Human Action Recognition System which searches and tracks human actions. To identify basic gesture/actions of the target (human) like sitting, standing, jumping, motionless etc. To develop basic situational awareness of the area of observation and identify behaviour of the subject based on data collected.

   ii. Work done: A novel HAR algorithm developed which uses both interest point based features in the form of good features to track as well as motion based features in the form of optical flow for feature detection within image sequences. The computed feature vectors are classified using a multilayer perceptron (MLP) network. The use of multiple features for motion descriptors enhances the quality of tracking. Resilient backpropagation algorithm is used for training the feedforward neural network reducing the learning time. The overall system accuracy is improved by optimizing the various parameters of the multilayer perceptron network.

   iii. Results achieved from the work (Give details of the papers and names of the

journals in which it has been published or accepted for publication and also about project competition won): Preparing Manuscript for submission for publication of the novel algorithm and basic implementation on Odroid XU4 SBC, in a few IEEE conferences on Computing, Signal Processing and IoT, etc.

iv. Has all the objectives been achieved as per plan. If not, state reasons: The overall algorithm for the proposed HAR system along with its subcomponents has been established. Moreover, the results up till feature extraction have been successfully implemented on Python. The Machine Learning algorithm for the feedforward neural network has been defined as well as the neural network function on OpenCV and Python established.

v. Please indicate the technical difficulties, if any, experienced in implementing the project: Significant challenges were faced during the Machine Learning and training/classification phase of the project due to the variability in the number of nodes, layers, as well as training of the feature vectors and passing them into the Multilayer Perceptron Network (Feed Forward Neural Network) as well as generation of the xml file. Moreover the challenges faced in the form of noise in the video background which leads to improper classification due to incorrect detection of features in high variance / dynamic background conditions, leading to a large offset in the correlated output vs the trained model. Moreover, additional challenges were faced during the installation of firmware on the Odroid XU4 module due to lack of community support.

Signature of Student 1

Jonti Talukdar

(14BEC057)

Signature of Student 2

Bhavana Mehta

(14BEC028)

Signature of Mentor

Dr. Tanish Zaveri

Professor,
Electronics and Communication Engineering,

Institute of Technology,

Nirma University, Ahmedabad.

Signature of Idea Lab Co-ordinator
Piyush Bhatasana
Idea Lab Co-ordinator,
Electronics and Communication,
Institute of Technology,
Nirma University, Ahmedabad.

Signature of Section Head
Dr. Dilip Kothari
Section Head,
Electronics and Communication,
Institute of Technology,
Nirma University, Ahmedabad.

Signature of HOD
Dr. Dilip Kothari
Electronics and Communication,
Institute of Technology,
Nirma University, Ahmedabad

Signature of Dr. Ankit Thakkar
Dr. Ankit Thakkar
Idea Lab Co-ordinator,
Institute of Technology,
Nirma University, Ahmedabad.

Signature of Director
Dr. Alka Mahajan,
Director,
Institute of Technology,
Nirma University, Ahmedabad

Contents

1.1 Introduction

Human Action Recognition is the process of recognizing various actions that people perform. These actions maybe walking, running, jumping, clapping, dancing etc. There are various approaches to implement Human Action Recognition on a system like skeleton representation, template matching, and use of Kinect for mapping movements. These different approaches have their own advantages and disadvantages which includes accuracy, easy implementation, environment condition, real time or non-real time and so on.

Human Action recognition (HAR) system aims to identify and recognize human actions by making a series of intelligent observations about the subject (human) using an array of sensors and intelligent computing algorithms. HAR is based on vision-based activity recognition which uses a camera to identify and observe the subject and based on processing of the video-feed, identify the action being performed by the subject. Basic actions include sitting, standing, walking, running, jumping, motionless etc. HAR from videos finds immediate applications in various fields like security, medicine, psychometric testing, smart cities, search and rescue, sign language recognition, sixth sense technology etc. HAR aims to use low cost open source technology to develop a HAR system, making it easier to use such intelligent systems in making everyday lives easier, safer, intelligent and increasingly connected.

This system can also be integrated into wide area network of other smart devices as well as existing networks thus facilitating its use on a much wider scale and geography at the same price. This project can also act as a backbone (sub-system) for other intelligent systems, like robots, rovers etc making it an incubator for developing even more intelligent & smarter technology which is connected, energy efficient and ultimately help in making this world a better place

1.2 Literature Survey

Research in this area has led to development of many new models and techniques in computer vision, which find applications in other spheres of life. For an algorithm to succeed, the method used for action representation and classification are of utmost importance. This has motivated the use of a plethora of different techniques such as local motion models, silhouette based models, SIFT, optical flow and classifiers like Bayes, SVM, KNN, ANN etc.

The overall objective of a HAR system is the automated analysis and interpretation of ongoing events and their context from video data [4]. Thus, the overall performance of the HAR system narrows down to the proper selection of feature vectors for action classification. Poppe [5] has divided image feature representations for HAR into two parts, global representations and local representations. The former is faster to compute and encodes the visual information of the whole image frame as feature vectors while the latter is more robust but computationally intensive and identifies specific patches of local activity around temporal interest points.

The use of silhouettes for tracking motion between multiple frames was pioneered by Bobick and Davis [6]. They used aggregate differences in video frames due to the dynamic nature of video data and generated a scalar field consisting of pixels whose intensities were a function of recent motion. This is called motion history image (MHI). Schüldt et al. [1] and Laptev et al. [7] described the use of representing motions in the form of local space-time interest points. A scale-space representation of the

image sequence is first obtained by using a Gaussian convolution kernel. Local features are then detected by computing the second moment matrix of the scale space representation within a Gaussian neighborhood of each point respectively. The local feature points obtained are classified using support vector machine (SVM). The results obtained on the test set are robust to variations in environment as well as noise.

An unsupervised learning approach was highlighted by Niebles et al. [8] in which Space-time interest points [2] are evaluated for local regions within the overall image sequence. Recognition is done by performing feature extraction on the input and then performing latent topic discovery through two models, probabilistic latent semantic analysis [9] and latent dirichlet allocation [10]. Recently, other classification techniques like boosting which combine several weak classifiers to form a stronger one have been proposed by Zhang et al. [11].

Since all these approaches used local descriptors, the results though highly accurate, could not be achieved in real time leading to significant delay. Hence, we have used a global approach which uses the strongest features in the video sequence in conjunction with the iterative optical flow algorithm to track and classify human actions. The key areas of our proposed HAR model can be quantified by the diagram shown below, Fig. 1.

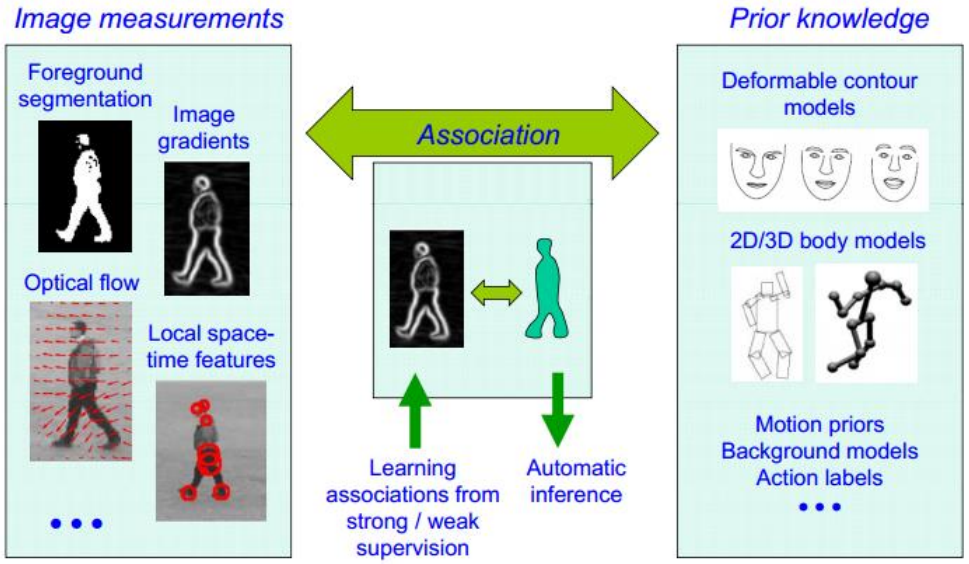

Fig 1: Key components of action recognition [4]

1.3 Major Objectives Proposed

The goal of the Action Recognition System is an automated analysis or interpretation of ongoing events and their context from video data. The main focus is to develop a novel and effective yet robust algorithm to recognize human actions and classify them in real time. The choice of approach to implement this application depends on the factors like processing time which can be deemed as minimal to obtain the required output in real time, size of neural network to reduce training time, length or size of feature vectors, which are to be computed without significant overhead in computing resources as well as the size of the training data.

The additional objective involves exploring the feasibility of implementing the given system on a Single Board Computing platform, Odroid XU4 to increase portability as well as flexibility of the operation of the on a real time basis.

1.4 Objectives Achieved

An overall algorithm for the overall implementation of the Human Action Recognition System has been successfully developed. The algorithm uses a combination of the iterative optical flow algorithm along with 'good features' [3] as the primary motion descriptors. The use of this approach ensures that multiple features are selected for motion description which ultimately enhances the quality of tracking of the subject of interest. Tracking of good features in dynamic video frames is also easier and computationally inexpensive as opposed to performing various transformations of entire frames in spatial and temporal domains as discussed in section 1.2. The overall algorithmic process can be highlighted in the given Fig. 2.

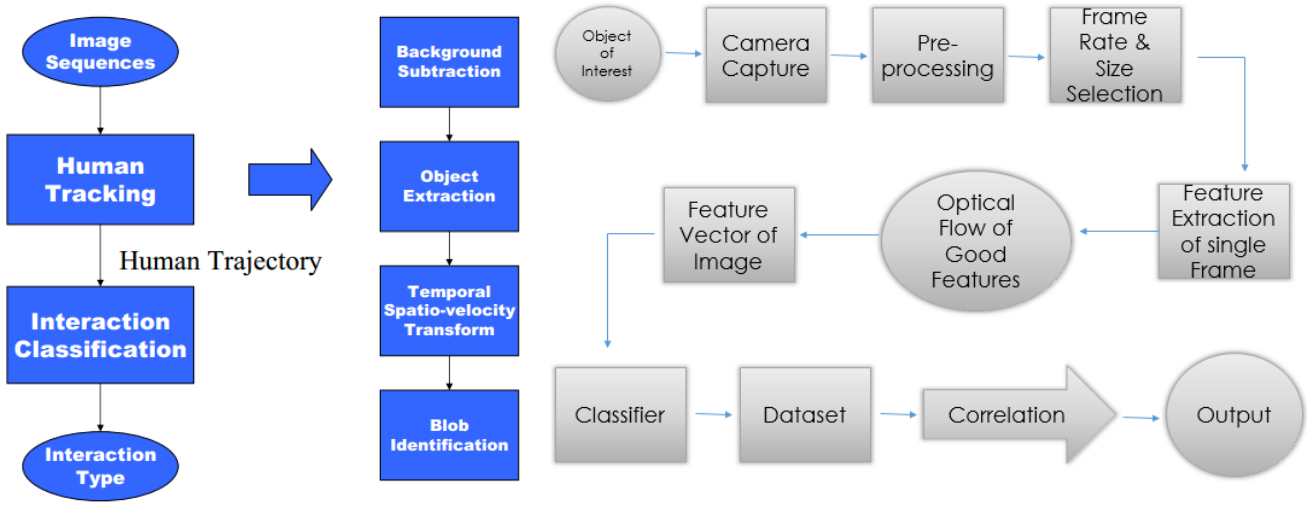

Fig 2: (a.)Traditional spatio temporal feature based algorithm, (b.) Our proposed Algorithmic Flow based on global features

Details about the algorithm can be further subdivided into the following sub steps:

1. Pre-processing: This stage involves the reduction in frame rate and size selection as well as conversion of the video frame from RGB to Grayscale in order to facilitate correlation with the training KTH data. It also involves the background subtraction results obtained during the process of subtracting consecutive frames to encapsulate the dynamic background.

2. Feature Extraction: The feature vectors are computed based on the good features [3], which is a modification of the corner detector algorithm [12] is evaluated on the pre-processed downsampled image sequences from the previous steps. The feature selection model as proposed by [3] is based on evaluating and monitoring interest point features whose selection maximizes the quality of tracking of the desired subject. Good features thus help in computing and tracking the *N* strongest features in a dynamic image sequence. Using motion based features which includes the Lukas Kanade pyramidal iterative optical flow algorithm [13], in conjunction with interest point features increases the accuracy of the system by many folds.

3. Classification: The extracted feature vectors are passed through a feedforward artificial neural network for classification. The Multilayer perceptron (MLP), which is used as the primary classifier, is composed of multiple nodal layers which are interconnected in a directed graph fashion. The resilient propagation (R-PROP) adaptive backpropagation algorithm [14-15] is used for training the MLP. As the model is trained, the weights of individual neurons are adapted locally based on their influence on an arbitrary error function. This direct adaptation of the weight updates reduces the learning steps significantly, is very efficient to compute in terms of storage and time and is also robust against the choice of starting values [15], thus improving the overall system performance.

Other objectives achieved include successful implementation up till the Feature Extraction stage as well as its testing on the KTH dataset, which has been used for training. Moreover, the proposed algorithm has been tested up till feature extraction stage onto the Raspberry Pi 3 Model B module as well.

1.5 Objectives Not Achieved

The Classification and Training phase is still under progress due to challenges which have been discussed in section 1.6.

1.6 Technical Difficulties Faced

The main difficulties faced in the design of the proposed Human Action Recognition system involves the Training and Classification stage. Training of the neural network although a significantly easy task, however requires an optimization in terms of the number of layers in the overall network as well as optimizing the size of the feature vector used for classification. However, the key challenge is encountered in the reduction of the same feature vectors as well as their identification on the real time video stream data, which is taken from the camera module.

The mismatch in the feature points detected on the input video stream with the background clutter reduces the overall accuracy of the proposed system and causes erroneous stream of data. Additionally, noise in the video background which leads to improper classification due to incorrect detection of features in high variance / dynamic background conditions is a key challenge, leading to a large offset in the correlated output vs the trained model. Moreover, additional challenges include the installation of firmware on the Odroid XU4 module due to lack of community support. Thus work is ongoing in figuring out a more robust yet highly directive way of identifying exclusive features of the human subject with respect to the cluttered background obtained from the real time video stream.

A possible contingency involves implementation of a smart segmentation algorithm on the input video sequence to detect only the moving subject and then applying the good feature model to more robustly identify the human subject followed by classification.

Additional challenges involve developing custom code for integration of the overall algorithm on the Embedded Linux platform / OS of the Odroid XU4 which does not fully support some of the libraries which run smoothly on the desktop Ubuntu environment.

***

1.7 Experimental Setup and Results

The overall experimental setup of the Single Board Platform is shown in Fig. 3.

The good features to track also known as Haaris Corner detector was implemented on python on video sequences for human action recognition and the results noted. Given below are the features which are being tracked. Simulation results attached in the figures 4-a and 4-b, below. Results of the Lukas Kanade Optical flow algorithm are also shown in Fig. 5 below.

Fig 3: Experimental Setup of the Odroid XU4 with the Wifi module and Camera Module Connected. It is running the standard Embedded Linux Ubuntu mate.

Fig 4: (a.) Good features extracted along with background subtraction results for clapping.

Fig 4: (a.) Good features extracted along with background subtraction results for running.

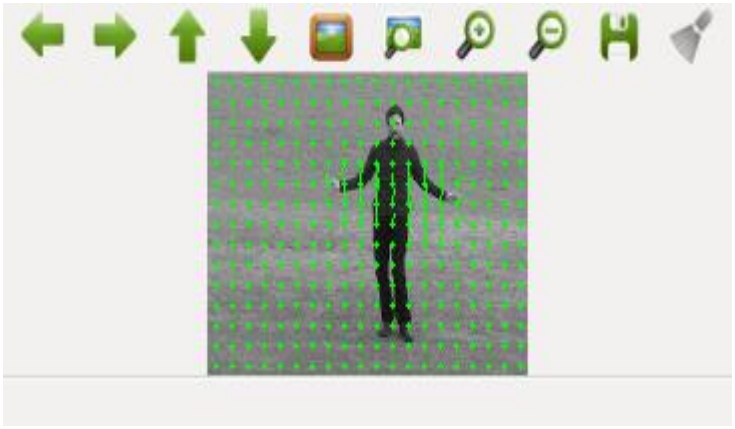

Fig 5: Implementation of the Lukas Kanade Iterative optical flow algorithm on the KTH datset test video.

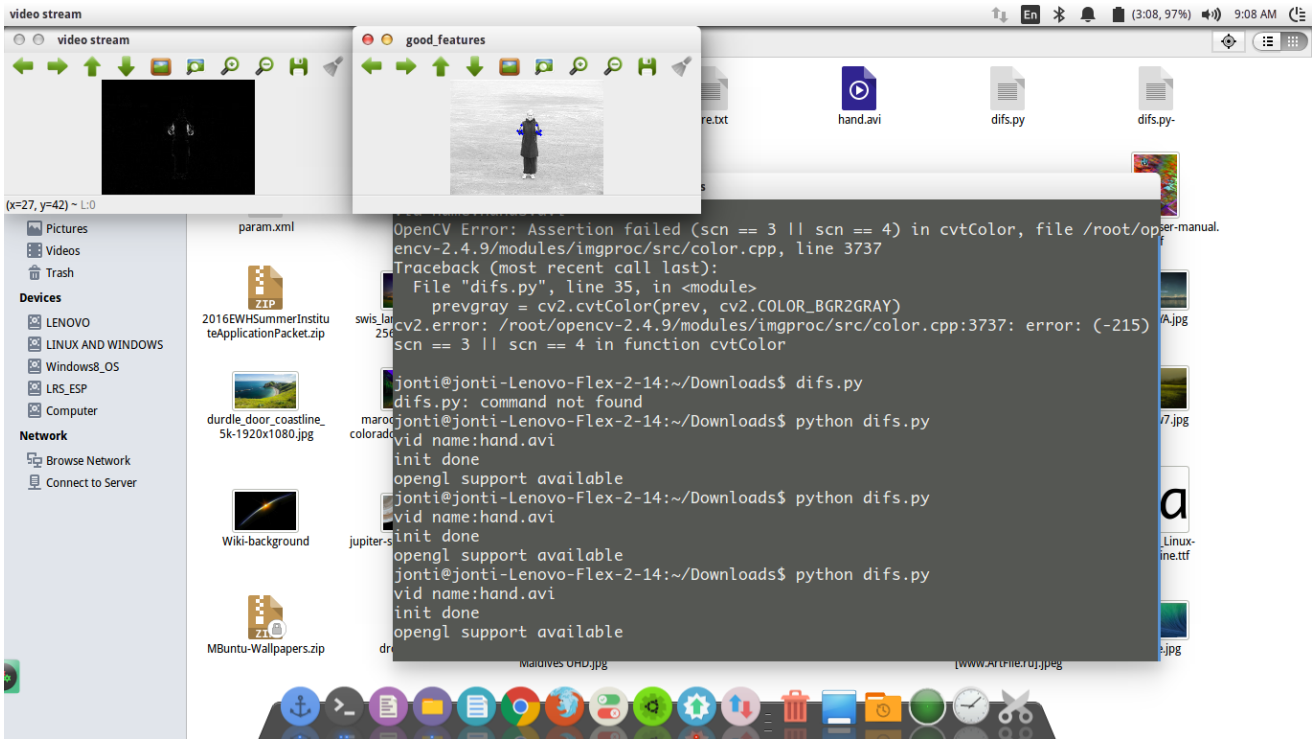

Fig 6: Frontal results obtained for classification of the clapping object class through the Python based classifier, identifying the key points being identified for a places of movement.

***

1.8 Budget Analysis

1. Budget Sanctioned: Rs. 17,000/-

2. Budget Utilized: Rs. 14,049/-

Sr .No	Item	Rate	Qty	Tax (5.5%)	Total
1	Odroid XU4	9,500	1	-	9,500
2	Odroid 720p camera	2,285	1	125.68	2,410.68
3	Odroid XU4 Case –base half	345.00	1	18.98	363.98
4	Odroid XU4 Case –top half	345.00	1	18.98	363.98
5	Odroid Wifi Module 3	1,195.00	1	65.73	1260.73
	Delivery Charge (DTDC)	150			150.00

3. Budget Unutilized: Rs. 2,950/-

***

1.9 Conclusion and Future Work

This report presents a conclusive treatment of the algorithmic approach used in Image Processing with the specific context of their implementation of HAR in Python, a high level language. Image Processing is a vibrant and diverse field with immediate applications in a wide variety of branches. Python being an easy and user friendly language provides a solid framework for the development of complex algorithms for processing images and a robust platform for processing large amount of data. This allows the developer to move from the mundane task of focusing on syntactical problems to the highly rewarding and important task of developing new models to integrate algorithms, develop systems and process data. The rise of Machine Learning has led to the growth of using such algorithms to automate image processing tasks which include automated recognition, detection and classification of images. This has also led to the growth of research and development of many new models and techniques in computer vision, which find applications in many spheres of life. The problems of image filtering, feature extraction, classification are discussed as well as the challenges observed are stated. With the growth already complete in the field of image processing, new trends still emerge and provide scope for potential development in the field of mathematical modelling for complex learning systems based on images. These include Human Action Recognition, Optical Character Recognition, Pattern Recognition and many more. Optimization in the field of Machine Learning can be done by focusing on developing new techniques to extract complex features easily and to classify large amount of data in a small number of instructions. Thus, with an increasing number of innovative ideas, this technology will also develop new solutions to pre-existing problems and help in making this world better & sustainable.

***

Bibliography

[1] C. Schuldt, I. Laptev and B. Caputo, "Recognizing human actions: a local SVM approach," In *Proc. IEEE International Conference on Pattern Recognition (ICPR)*, 2004, pp. 32-36.

[2] I. Laptev and T. Lindeberg, "Space-time interest points," In *Proc. IEEE International Conference on Computer Vision (ICCV)*, Nice, France, 2003, pp. 432-439.

[3] Jianbo Shi and C. Tomasi, "Good features to track," In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (ICVPR)*, Seattle, WA, 1994, pp. 593-600.

[4] K.Maithili, K. Rajeswari, R. Mohanapriya, D. Krithika, "An Efficient Human Action Recognition System Using Single Camera and Feature Points", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, No. 2, 2013.

[5] R. Poppe, "A survey on vision-based human action recognition," Image and Vision Computing, Vol. 28, No. 6, pp. 976–990, 2010.

[6] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," in IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 23, No. 3, pp. 257-267, 2001.

[7] I. Laptev and T. Lindeberg, "Local descriptors for spatiotemporal recognition" In *Proc. ECCV Workshop on Spatial Coherence for Visual Motion Analysis*, 2004, pp. 91-103.

[8] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words." International Journal of Computer Vision, Vol. 79, No. 3, pp. 299-318, 2006.

[9] T. Hofmann, "Unsupervised Learning by Probabilistic Latent Semantic Analysis," Machine Learning, Vol. 41, No. 2, pp. 177-196, 2001.

[10] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation." Journal of Machine Learning Research, Vol. 3, pp. 993–1022, 2003.

[11] Li, Z. Stan, L. Zhu, Z. Zhang, A. Blake, H. Zhang, and H. Shum, "Statistical learning of multi-view face detection." *European Conference on Computer Vision*. Springer Berlin Heidelberg, 2002, pp. 67-81

[12] C. Harris and M. Stephens, "A Combined Corner and Edge Detector," In *Proc. Alvey Vision Conference*, Vol. 15. No. 50, pp. 147-151, 1988.

[13] J.Y. Bouguet, "Pyramidal Implementation of the Lucas Kanade Feature Tracker Description of the Algorithm," Intel Corporation, Vol. 5, No. 1-10, 2001.

[14] Y. LeCun, L. Bottou, G.B. Orr, and K. Muller, "Efficient Backprop," Neural Networks, Tricks of the Trade, Springer, 1998.

[15] M. Riedmiller and H. Braun, "A direct adaptive method for faster backpropagation learning: the RPROP algorithm," *IEEE International Conference on Neural Networks*, San Francisco, CA, 1993, pp. 586-591.