Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 07, July-2019

Enhancement of Video Action Recognition using Semi-Supervised Learning

Saniya M Sunil
M.tech student
Department of Computer Science and Engineering,
LBS Institute of Technology for Women
Thiruvananthapuram, India

Dileep V K
Assistant Professor
Department of Computer Science and Engineering,
LBS Institute of Technology for Women
Thiruvananthapuram, India

***Abstract*—Many efforts has been put forward to achieve the performances of human activity or the gesture recognition in videos by the transformation of action knowledge fetched from the still images to videos. In this paper, an adaptation method has been used to transform activity or the action recognition in videos by adapting knowledge from images. The adapted knowledge is used to learn the mutual related semantic actions by inquiring the common elements of both labelled videos and images. The existing action recognition method use supervised method for action recognition so that it is very difficult to collect the labelled videos that cover different types of actions. In such situation, over fitting would be an inherent problem and the performance of activity recognition is confined and becomes excessively complicated. Thus, the over-fitting can be mitigated and the performance of human activity recognition is improved. Meanwhile, we expand the adaptation method to a semi-supervised framework which can both labelled and unlabeled videos.**

***Keywords— Semi-supervised learning, Action recognition, Knowledge adaptation, labelled and unlabelled videos*.**

I. INTRODUCTION

With the fast progress of Internet and mobile phone, activity acknowledgment in individual recordings has turned into an essential research point because of its wide applications such as automatic video tracking, video annotation, video explanation, and so on. Recordings which are uploaded on the web by the users are transferred by clients and created by handy cameras may contain extensive camera shake and disturbances, hindrance, and jumbled foundation. Therefore, these recordings contain huge intra class variations within the same category in this manner. Hence, altogether now it is a challenging job to recognize human actions in such similar videos. A large number of local or confined features, motion scale invariant feature transform are extracted from videos then and all local features are quantized into a histogram vector using bag-of-words illustration. Then the vector-based classifiers are finally used to perform the action recognition in testing videos. When the videos are entirely simple, recognition methods have accomplished promising results. However, noises and uncorrelated information may be incorporated into the bow during the extraction and quantization of the local features. Therefore, we can come to an idea that these methods are generally not robust and cannot be generalized well when the videos contain specific camera shake, occlusion, cluttered background and so on.

In order to attain recognition accuracy, meaningful elements of actions such as related objects, human gestures, behavior etc. should be applied to form a clearer semantic understanding of human actions. The effectiveness of leveraging related object or human poses or actions have been demonstrated in recent efforts. The methods may require to training process with huge amounts of videos to obtained good performance, especially for real world videos. Though, it is really challenging to collect enough labelled videos that cover a distinct range of action poses. Knowledge alteration or adjustment from images to videos have exhibited improved performance in application areas of cross media recognition and retrieval. Knowledge adaptation is also known as transfer learning in which the target is to disseminate the knowledge from ancillary domains to target domains.

II. RELATED WORK

In 2008, Christian Thurau, et.al proposed Pose Primitive Based Human Action Recognition in Videos or Still Images. In this paper, they have presented a pose based approach for action recognition from still images and image sequences. This approach does not involve any background subtraction or a still camera and can be certainly extended to multiple persons.

In 2015, Y. Han proposed Semi Supervised Features Selection via Spline Regression for Video Semantic Recognition which issued to enhance both the efficiency and accuracy of the video semantic action recognition, it can perform feature selection on the derived video features to select the subset of characters form the high dimensional feature set for a dense and exact video data representation. This discloses semi-supervised attribute selection algorithms to better recognize the appropriate video features, which are biasive target classes by effectively utilizing the information underlying the large quantity of unlabeled video data.

Kernelized multiview projection for robust action recognition has been intended by L. Shao, et.al in 2015. In this paper, they have introduced an efficient sub-space training framework based on KMP for human action or gesture recognition. KMP can encrypt a different kinds of features in different ways to attain a semantically significant embedding. A relevant feature of KMP is that it is able to effectively

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 07, July-2019

explore the equivalent property of distinct views and eventually detects a unique low-dimensional subspace where the distribution of each view is adequately smooth and differentiative.

In 2016, L. Liu, et.al proposed Learning Spatio-Temporal Representations For Action Recognition: A Genetic Programming Approach on the Genetic programmingapproach. In this article, instead of applying handmade features, we instinctively learn the spatial as well as temporal motion characters or features for action recognition. This is used to achieve via a generative evolutionary technique, that means the Genetic Programming, an automatic method which derives the motion or the gesture feature descriptor structure on a community of primitive or elementary 3-D operators.

In 2016, M. Yu, et.al has been worked on the Structure Preserving Binary Representations for RGB-D Action Recognition which aims on the Local representation for RGB-D (which is generally a combination of RGB image and its depth information) video data fusion with a structure or an arrangement preserving projection.

In 2016, B. Ma, et.al proposed a paper based on Discriminative Tracking Using Tensor Pooling to represent target templates and candidates directly with sparse coding tensor. Local sparse representation has been successfully applied to visual tracking, owing to its discriminative nature and robustness against local noise and partial occlusions. Local sparse codes computed with a template actually constitute a three-order tensor according to their original layout, although most existing pooling operators convert the codes to a vector by concatenating or computing where it is used to deliver more informative and structured information, which potentially enhances the discriminative power of the appearances model and improves the tracking performances.

In 2016, C. Li, et. al has been worked on Transfer Latent SVM for joint Recognition and Localization of Actions in videos and it is based on web images and weakly annotated training videos. The model takes training videos which are only annotated with action label as input for alleviating the laborious and time-consuming manual annotations of action locations. For the purpose of improving the localization we collect an number of web images which are annotated with both action labels and action location to learn a discriminative model by enforcing the local similarities between videos and web images.

Multi Surface Analysis for Human Action Recognition in Video has been presented by Hong-Bo Zhang, et. al in 2016. They proposed a novel multi-surface feature named 3SMF. The prior probability is estimated by an SVM, and the posterior probability is computed by the NBNN algorithm with STIP. We model the relationship score between each video and action as a probability inference to bridge the feature descriptors and action categories.

The Comparison between the SIFT and SURF has been proposed by Darshana Mistry and Asian Banerjee in 2017. SIFT is used for finding uniqueness features. In this paper it is said that the SURF is three times better than that of SIFT because of using the integral image and box filters. Here the SIFT will take more time to extract the features when compared to SURF. SIFT and SURF, both are robust method inorder to find feature detection and matching.

In 2017, B. Ma, L. Huang, J. Shen and L. Shao proposed a paper based Label Information Guided Graph Construction For Semi-Supervised Learning which use the label information of observe sample in the label propagation stage, while ignoring such valuable information when learning the graph. The enforcing the weight of edges between labeled samples of different classes to the state of the art graph learning methods, such as the low rank representation learning method called semi supervised low rank representation.

III. PROPOSED METHODOLOGY

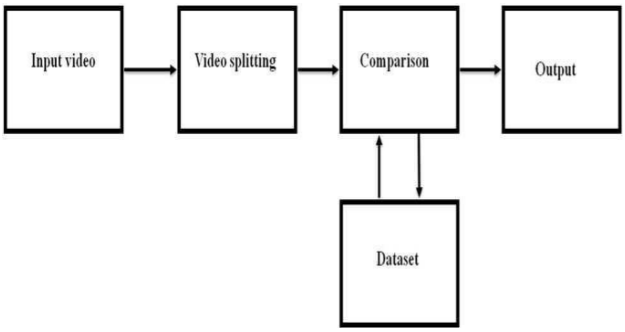

Fig 1. Block diagram

A) Key frame extraction

The main steps of the key frames extraction include the following. First, the color histogram of every five frames is calculated. Second, the histogram is subtracted with that of the previous frame. Third, the frame is a shot boundary if the subtracted value is larger than an empirically set threshold. Once we get the shot, the frame in the middle of the shot is considered as a key frame.

IV. SEMI-SUPERVISED LEARNING

Semi-supervised learning is a family of machine learning tasks and techniques which make use of unlabelled data for training low amount of unlabeled data. Semi-supervised learning falls among unsupervised learning as well as supervised learning.

IVA: The proposed semi-supervised knowledge adaptation method is actually designed for adapting knowledge from images based on the feature A and for utilizing the videos as the target domain based on feature AB.

V. DATASET

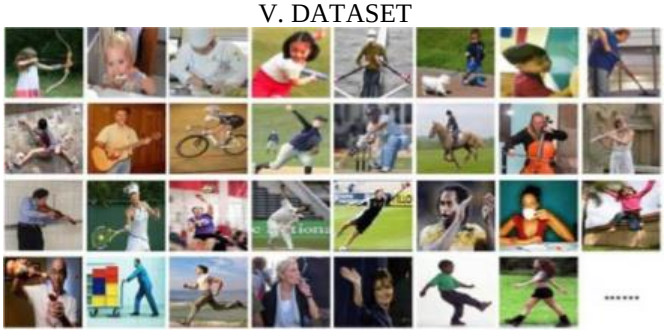

Fig 2. Sample images from dataset

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 07, July-2019

VI. RESULTS

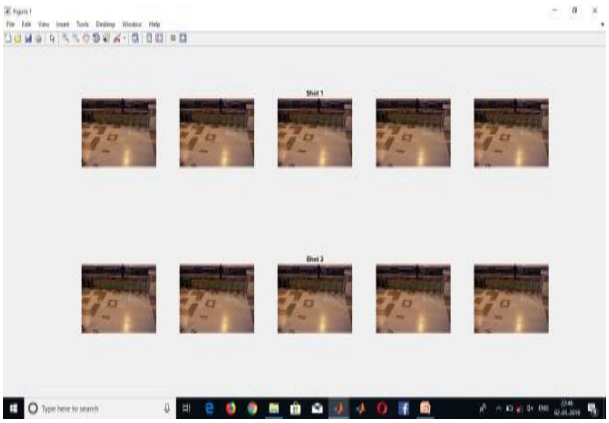

Fig 3. Extraction of frames from the selected video

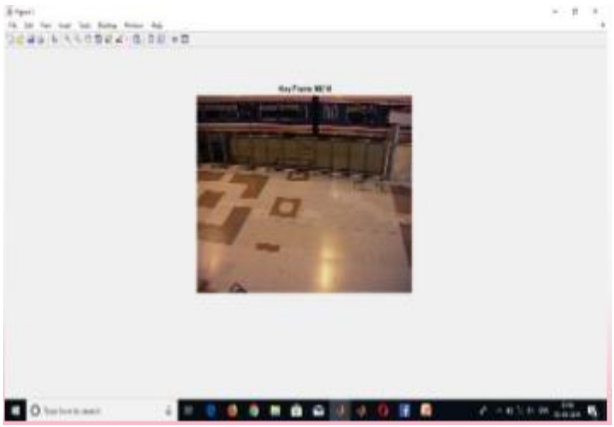

Fig 4. Key frame extraction from the shot

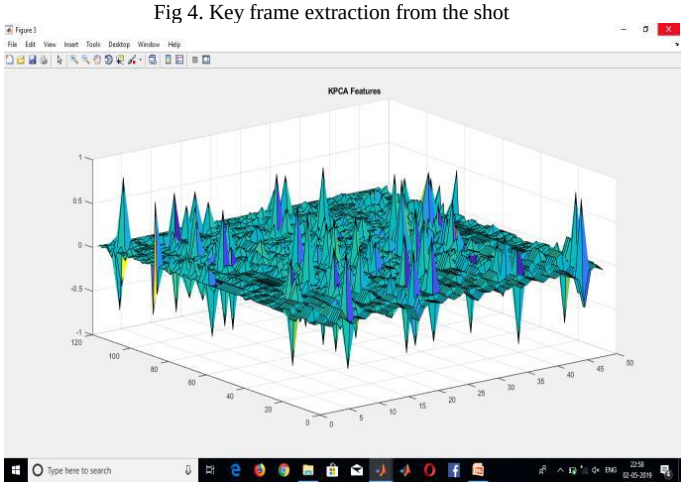

Fig 5. Mapping the combination of video and image features by KPCA method

VII. CONCLUSION

To achieve the overall performance of action recognition, we propose a classifier of Image to Video Adaptation, which is able to acquire the knowledge from images based on commonly visual features. At the same time, it can completely utilize the heterogeneous features of unlabeled videos for improving the performance of human action recognition in the videos. Empirical results reveal that the knowledge learned from the images can make an impact in the recognition accuracy or fidelity of the videos. Moreover, the results prove that the intended IVA(Image to video adaptation) exhibit the improved performances of human action recognition.

REFERENCES

[1] Christian Thurau, Vaclav hlavac, “Pose primitive based human action recognition in videos or still images”, Proceedings of the conference of computer vision and pattern recognition, Anchorage, Alaska, USA, June 2008.

[2] Yahong Han, Yi Yang, Zhigang Ma, Yan Yan, Nicu Sebe, Xiaofang Zhou, “Semi-Supervised Feature Selection via Spline Regression for Video Semantic Recognition”, MANUSCRIPT SUBMITTED TO IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, February 2015.

[3] L. Shao, L. Liu, and M. Yu, “Kernelized multiview projection for robust action recognition”, International Journal of Computer Vision, October 2015.

[4] Li Liu, Ling Shao, Xuelong Li, and Ke Lu, “Learning Spatio-Temporal Representations for Action Recognition: A Genetic Programming Approach”, IEEE TRANSACTIONS ON CYBERNETICS, VOL. 46, NO. 1, January 2016.

[5] Mengyang Yu, Li Liu, and Ling Shao, “Structure-Preserving Binary Representations for RGB-D Action Recognition”, IEEE transactions on pattern analysis & machine intelligence, VOL. 38, NO. 8, August 2016.

[6] Bo Ma, Lianghua Huang, Jianbing Shen, and Ling Shao, “Discriminative Tracking Using Tensor Pooling”, IEEE TRANSACTIONS ON CYBERNETICS, VOL. 46, NO. 11, November 2016.

[7] Cuiwei Liu, Xinxiao Wu, and Yunde Jia, “Transfer Latent SVM for Joint Recognition and Localization of Actions in Videos”, IEEE TRANSACTIONS ON CYBERNETICS, VOL. 46, NO. 11, November 2016.

[8] Hong-Bo Zhang, Qing Lei, Bi-Neng Zhong, Ji-Xiang Du, Jialin Peng, Tsung-Chih Hsiao and Duan-Sheng Chen, “Multi-surface analysis for human action recognition in video”, Zhang et al. SpringerPlus, 2016.

[9] Darshana Mistry and Asim Banerjee, “Comparison of Feature Detection and Matching Approaches: SIFT and SURF”, GRD Journals- Global Research and Development Journal for Engineering, Volume 2, Issue 4, March 2017.

[10] Liansheng Zhuang, Zihan Zhou, Shenghua Gao, Jingwen Yin, Zhouchen Lin and Yi Ma, “Label Information Guided Graph Construction for Semi-Supervised Learning”, IEEE transactions on image processing, VOL. 26, NO. 9, September 2017.