A SURVEY ON THE ENHANCEMENT OF VIDEO ACTION RECOGNITION USING SEMI-SUPERVISED LEARNING

Saniya M Sunil[1], Dileep V K[2]

[1]Saniya M Sunil, M.Tech Student, Department of Computer Science and Engineering, LBS Institute Of technology For Women, Kerala, India
[2]Dileep V K , Assistant Professor, Department of Computer Science and Engineering, LBS Institute Of technology For Women, Kerala, 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 overfitting 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.**

1. 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.

2. ACTION RECOGNITION ENHANCEMENT METHODS

A. Pose Primitive Based Human Action Recognition in Videos or Still Images.

In 2008, Christian Thurau, et. al proposed this paper. 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 the learning mode, parameter representing poses and actions are evaluated from videos. On the other hand, the method can be used for both videos and still images in the run mode. For recognizing pose, they used Histogram of Oriented Gradient (HOG) based descriptor to better contend with attached poses and cluttered or disordered background. Actions are represented using HOG and the action recognition is done based on a simple histogram comparison. The proposed approach does not confide on background subtraction or dynamic features. Therefore allows for action recognition in still images.

B. Evaluating Color Descriptors for Object and Scene Recognition

K. E. Van De Sande, et. al presented this paper in 2010. The paper literally describes the invariance or consistent properties and the discreteness of color descriptors in a ordered way. The invariable properties of color descriptors are exhibited logically using a taxonomy or allocation based on invariable properties with respect to the photometric transformations. The discreteness of color descriptors can be assessed analytically using two benchmarks from the image domain as well as the video domain. From the hypothetical results, any changes in the light intensity, the advantage of invariance is category-specific.

C. Grouplet: A structured image representation for recognizing human and object interactions.

In 2010, B. Yao et. al propounded this paper. They presented a novel image feature representation concept called the "grouplet". Grouplet defines a small group which captures the structured or a disciplined information of an image by encoding or encrypting a number of differentiative visual characters and their spatial configurations. By applying a dataset of seven different people-playing-musical-instrument (PPMI) activities, they demonstrated that the grouplets are more efficient for the classification and detection of human-object interactions compared to the other state-of-the-art methods. For example, consider an example people-playing-musical-instrument (PPMI); that distinguishes an individual playing violin from an individual holding a violin needs elegant prominence of peculiar image features and the feature arrangements that distinguish these two scenes.

D. Multimedia Retrieval Framework Based on Semi-Supervised Ranking And Relevance Feedback.

In 2012 Y. Yanghas has been worked on this paper and the paper describes two types of algorithm which is a local regression and global alignment algorithm to learn a robust Laplacian matrix for ranking or grading. The local linear regression method is used to anticipate the ranking scores of its neighboring points whereas the semi supervised are used for long term relevance feedback(RF) algorithm to reline the multimedia input data representation in the multimedia attribute space and the history relevance feedback information provided by users. And they are applied to various content-based multimedia retrieval application which include cross media retrieval and 3D motion or pose data retrieval. It is used on four data sets have demonstrated its advantage in preciseness, robustness, flexibility, and computational efficiency. The proposed framework attains noteworthy performance.

E. Visual event recognition in videos by learning from Web data.

This paper has been worked by L. Duan, et. al in 2012. They have presented a visual event recognition framework which can be used for the consumer videos by authorizing a huge quantity of loosely or freely labeled web videos such as from YouTube website. Though the event recognition is extremely a challenging computer vision task, it has attracted growing attention because of the increasingly critical demand on and recovering from a huge amount of unconfined consumer videos. A new pyramid matching method called ASTPM which means the Aligned-Space-Time-Pyramid-Matching algorithm and a novel based transfer learning method which is referred to as A-MKL is depicted here to be better fuse the information from multiple pyramid levels and the different types of confined features.

F. Dense Trajectories and Motion Boundary Descriptors for Action Recognition.

H. Wang, et. al propounded this paper in 2013. This paper defines the video illustration which are based on the dense trajectories and the motion boundary descriptors. Trajectories(path) can capture or it will get the local gesture information or facts of the videos. A dense representation assures a proper handling of foreground motion and also the surrounding context. The state-of-art optical flow algorithm provides a robust and effective extraction of the dense trajectories. illustration and also could be applied in the circumstances of action localization and the video retrieval.

G. Learning Discriminative Key Poses for Action Recognition.

In 2013, L. Liu, et. al has been worked on this paper. A new method have been presented based on the key poses representation to represent The poses in the video frames are distinguished by the proposed extensive pyramidal features which can include the feature maps like Gabon, Gaussian, and wavelet pyramids. All these features are used to encode the orientation, intensity and configuration information and thereby providing an informative representation of the human poses or actions. Here to learn the subset of various discriminative poses for representing actions, we utilize the Ada boost algorithm.

H. Action recognition using non-negative action component representation and sparse basis selection.

In 2014, H. Wang proposed this paper by using the higher level action units for representing the human actions in videos and then depending upon those units, a novel sparse based model has been developed for the human action recognition. In this method, they have been introduced three interconnected components. At first, they has been presented a new context aware spatial and a temporal descriptor for enhancing the bias of the locally spatial as well as the temporal descriptors which was applied traditionally. Secondly, they attempted to study understand or the action units by using the graph equalized positive matrix factorization from the statistics of the context-aware descriptors. This will leads to the part-based illustration and also encodes or conceals the geometrical information. These units effectively connect the acceptable gap in human action recognition. Thirdly, we defines a sparse model based on a joint L2,1-norm form which has been proposed to perform the feature selection to preserve the representative components and conceal the noise in the action units.

I. Learning with augmented features for heterogeneous domain adaptation.

L. Duan, et. proposed this paper in 2014. They have been introduced a modern technique for HDA which means Heterogeneous-Domain-Adaptation. In HFA, they augmented the heterogeneous features from the source as well as the target domains by using two newly proposed feature mapping functions. They first proposed to find the two projection matrices for the source as well as the target data by applying the standard learning algorithm Support Vector Machine with the hinge loss in both linear and the nonlinear circumstances.

J. Knowledge adaptation with partially shared features for event detection using few exemplars.

In 2014, Z. Ma presented this paper. They introduced the research exploration of Multimedia event detection along with few exemplars. Since this focuses on more generic, complicated and meaningful events, this is an important research issue that reflects in our daily activities. In addition, the situation we have faced in the real-world needs that only few positive examples are used. For achieving great performance, we have proposed to acquire strength from the available concepts-based videos for the MED with little exemplars.

K. Multilevel Chinese Takeaway Process and Label based Process For Rule Induction in the context of Automated Sports Video Annotation.

In 2014, A. Khan, et. al worked this paper and the paper propose four alternatives of a hierarchical hidden Markov model strategy for indication in the frame of reference of automated sport video annotation comprising a Multilevel-Chinese-Takeaway; which is a process and a recent technique known as the Cartesian-product-label-based-hierarchical bottom-up clustering; which is called CLHBC method that utilize previous information incorporated within the label design. Optimal performance is acquired using the hybrid topological arrangement with CLHBC generated case labels.

L. Semi Supervised Features Selection via Spline Regression for Video Semantic Recognition.

In 2015, Y. Han proposed a paper 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.

M. Kernelized multiview projection for robust action recognition.

This method 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 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.

N. Learning Spatio-Temporal Representations For Action Recognition: A Genetic Programming Approach.

In 2016, L. Liu, et. al proposed an article based on the Genetic programming approach. 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. The GP evolving feature selection or derivation methods is estimated on four common action datasets, particularly KTH, HMDB51,UCF. This evolves an adaptive learning style using GP (Genetic Programming) to evolve biasive spatio-temporal illustration, which all together fuse the color and motion data or knowledge for high-level action recognition tasks.

O. Structure Preserving Binary Representations for RGB-D Action Recognition.

In 2016, M. Yu, et. al has been worked on this paper 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. For obtaining a general characteristic for the video data, we convert the problem to defining the gradient fields of RGB and depth information of video sequences. With the local click alterations or the changes of the gradient fields, which comprise the orientation and magnitude of the neighborhood of each point, a new kind of continuous local descriptor called local flux feature is acquired. This will obtain a fused local a binary illustration for RGB-D human action recognition.

P. Discriminative Tracking Using Tensor Pooling.

In 2016, B. Ma, et. al proposed a paper 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.

Q. Transfer Latent SVM for joint Recognition and Localization of Actions in videos.

In 2016, C. Li, et. al has been worked on this paper 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.

R. Multi Surface Analysis for Human Action Recognition in Video.

This paper 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 main contributions of our study is that a new holistic feature (3SMF) is proposed to represent video. 3SMF can reflect the difference of the action in different surfaces.

S. The Comparison between the SIFT and SURF.

The paper 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.

T. Label Information Guided Graph Construction For Semi-Supervised Learning.

In 2017, B. Ma, L. Huang, J. Shen and L. Shao proposed a paper based on semi-supervised learning methods 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.

3. 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] K. E. Van De Sande, T.Gevers, and C. G. Snoek, "Evaluating color descriptors for object and scene recognition," IEEE Trans. Pattern Analysis, vol. 32, no. 9, pp. 1582–1596, Sep. 2010.

[3] Bangpeng Yao, Li Fei-Fei, "Grouplet: A structured image representation for recognizing human and object interactions", IEEE Computer Society

Conference on Computer Vision and Pattern Recognition, 2010.

[4] Yi Yang, Feiping Nie, Dong Xu, Jiebo Luo, Yueting Zhuang, and Yunhe Pan, "A Multimedia Retrieval Framework Based on Semi-Supervised Ranking and Relevance Feedback", IEEE transactions on pattern analysis & machine intelligence, vol 34, No.4, April 2012.

[5] Lixin Duan, Dong Xu, Ivor Wai-Hung Tsang, and Jiebo Luo, "Visual Event Recognition in Videos by Learning from Web Data", IEEE transactions on pattern analysis & machine intelligence, VOL. 34, NO. 9, SEPTEMBER 2012.

[6] Heng Wang, Alexander Klaser, Cordelia Schmid, Cheng-Lin Liu,, "Dense Trajectories and Motion Boundary Descriptors for Action Recognition", International Journal of Computer Vision, Springer Verlag, 2013.

[7] Li Liu, Ling Shao, Xiantong Zhen, and Xuelong Li, "Learning Discriminative Key Poses for Action Recognition", IEEE TRANSACTIONS ON CYBERNETICS, VOL. 43, NO. 6, December 2013.

[8] Haoran Wang, Chunfeng Yuan, Weiming Hu, Haibin Ling, Wankou Yang, and Changyin Sun, "Action Recognition Using Non-negative Action Component Representation and Sparse Basis Selection", IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 23, NO. 2, February 2014.

[9] Lixin Duan, Dong Xu, and Ivor W. Tsang, "Learning with Augmented Features for Heterogeneous Domain Adaptation", IEEE transactions on pattern analysis & machine intelligence, VOL. 36, NO. 6, June 2014.

[10] Zhigang Ma, Yi Yang, Nicu Sebe, and Alexander G. Hauptmann, "Knowledge Adaptation with Partially Shared Features for Event Detection Using Few Exemplars", IEEE transactions on pattern analysis & machine intelligence, VOL. 36, NO. 9, September 2014.

[11] A. Khan, D. Windridge, and J. Kittler, "Multi-Level Chinese Takeaway Process and Label-Based Processes for Rule Induction in the Context of Automated Sports Video Annotation", IEEE TRANSACTIONS ON CYBERNETICS, October 2014.

[12] 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.

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

[14] 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.

[15] 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.

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

[17] 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.

[18] 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.

[19] 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.

[20] 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.