© 2019 IJSRSET | Volume 6 | Issue 2 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Engineering and Technology
DOI : https://doi.org/10.32628/IJSRSET1961166

Robust Feature-Based Automated Multi Viewhuman Action Recognition System Using Machine Learning

Paruchuri Yogesh[1], R. Dillibabu[1], Palaniappan P[1], Prof. L. Ashok Kumar[2], Prof. Navarajan[2]

[1]Department of ECE, Panimalar Institute of Technology, Poonamallee, Chennai, Tamil Nadu, India

[2]Assistant Professor, Department of ECE, Panimalar Institute of Technology, Poonamallee, Chennai, Tamil Nadu, India

ABSTRACT

Automated human action Recognition has the potential to play an important role in Public security. In this project it compares three practical, reliable and generics systems for multiview video based human action recognition namely the nearest classifier, Gaussian mixture model classifier and nearest mean classifier.

Keywords : MoSIF, BoWs, STIP, EMD, SVM, LST Feature, BI-Linear Interpolation , Classifier, K-Nearest Neighbour, Feature Extraction

I. INTRODUCTION

With the speedy advance of net and sensible phone, action recognition in personal videos made by users has become a vital analysis topic because of its wide applications, like automatic video chase and video annotation , etc. shopper videos on the online are uploaded by users and made by hand-held cameras or sensible phones, which can contain respectable camera shake, occlusion, and littered background. Thus, these videos contain massive intra category variations inside the identical linguistics class. it's currently a difficult task to acknowledge human actions in such videos. several action recognition strategies followed the traditional framework. First, an outsized range of native motion options (e.g., coordinate system interest points (STIP) motion scale invariant feature remodel (MoSIFT) etc.) are extracted from videos. Then, all native options are measure into a bar graph vector mistreatment bag-of-words (BoWs) illustration. Finally, the vector-based classifiers (e.g., support vector machine [9]) are wont to perform recognition in testing videos. once the videos are straightforward, these action recognition strategies have achieved promising results. However, noises and unrelated data could also be incorporated into the throughout the extraction and division of the native options. Therefore, these strategies are typically not strong and will not be generalized well once the videos contain respectable camera shake, occlusion, littered background, and so on. so as to enhance the popularity accuracy, significant parts of actions, e.g., connected objects, human look, posture, and so on, ought to be used to make a clearer linguistics interpretation of human actions. Recent efforts have incontestable the effectiveness of leverage connected objects or human poses.

Objective :

Currently, most of the knowledge adaptation algorithms require sufficient labeled data in the target domain. In real world applications, however, most videos are unlabeled or weak-labeled. Collecting well-labeled videos is time consuming and labor intensive. Simultaneously utilizing labeled and unlabeled data is beneficial for video action recognition. In order to enhance the performance of

action recognition, we explore how to utilize semi-supervised learning to leverage unlabeled data and thus to learn a more accurate classifier.

II. LITERATURE SURVEY

1. **MoSIFT: Recognizing Human Actions in Surveillance Videos CMU-CS-09-161 Ming-yu Chen and Alex Hauptmann:**

   The goal of this paper is to build robust human action recognition for real world surveillance videos. Local spatio-temporal features around interest points provide compact but descriptive representations for video analysis and motion recognition. Current approaches tend to extend spatial descriptions by adding a temporal component for the appearance descriptor, which only implicitly captures motion information. We propose an algorithm called MoSIFT, which detects interest points and encodes not only their local appearance but also explicitly models local motion. The idea is to detect distinctive local features through local appearance and motion. We construct MoSIFT feature descriptors in the spirit of the well-known SIFT descriptors to be robust to small deformations through grid aggregation. We also introduce a bigram model to construct a correlation between local features to capture the more global structure of actions. The method advances the state of the art result on the KTH dataset to an accuracy of 95.8%. We also applied our approach to 100 hours of surveillance data as part of the TRECVID Event Detection task with very promising results on recognizing human actions in the real world surveillance videos.

2. **Observing Human-Object Interactions: Using Spatial and Functional Compatibility for Recognition Abhinav Gupta, Member, IEEE, AniruddhaKembhavi, Member, IEEE, and Larry S. Davis, Fellow, IEEE**

   Interpretation of images and videos containing humans interacting with different objects is a daunting task. It involves understanding scene/event, analyzing human movements, recognizing manipulable objects, and observing the effect of the human movement on those objects. While each of these perceptual tasks can be conducted independently, recognition rate improves when interactions between them are considered. Motivated by psychological studies of human perception, we present a Bayesian approach which integrates various perceptual tasks involved in understanding human-object interactions. Previous approaches to object and action recognition rely on static shape/appearance feature matching and motion analysis, respectively. Our approach goes beyond these traditional approaches and applies spatial and functional constraints on each of the perceptual elements for coherent semantic interpretation. Such constraints allow us to recognize objects and actions when the appearances are not discriminative enough. We also demonstrate the use of such constraints in recognition of actions from static images without using any motion information.

3. **Visual Event Recognition in Videos by Learning from Web Data LixinDuan Dong Xu Ivor W. Tsang School of Computer Engineering Nanyang Technological University {S080003, DongXu, IvorTsang}@ntu.edu.sg JieboLuo Kodak Research Labs Eastman Kodak Company, Rochester, NY, USA**

   We propose a visual event recognition framework for consumer domain videos by leveraging a large amount of loosely labeled web videos (e.g., from YouTube). First, we propose a new aligned space-time pyramid matching method to measure the distances between two video clips, where each video clip is divided into space-time volumes over multiple levels. We calculate the pair-wise distances between any two volumes and further integrate the information from different volumes with Integer-flow Earth Mover's Distance (EMD)

to explicitly align the volumes. Second, we propose a new cross-domain learning method in order to 1) fuse the information from multiple pyramid levels and features (i.e., space-time feature and static SIFT feature) and 2) cope with the considerable variation in feature distributions between videos from two domains (i.e., web domain and consumer domain). For each pyramid level and each type of local features, we train a set of SVM classi- fiers based on the combined training set from two domains using multiple base kernels of different kernel types and parameters, which are fused with equal weights to obtain an average classifier. Finally, we propose a cross-domain learning method, referred to as Adaptive Multiple Kernel Learning (A-MKL), to learn an adapted classifier based on multiple base kernels and the prelearned average classi- fiers by minimizing both the structural risk functional and the mismatch between data distributions from two domains. Extensive experiments demonstrate the effectiveness of our proposed framework that requires only a small number of labeled consumer videos by leveraging web data.

4. **Action Recognition Using Nonnegative Action Component Representation and Sparse Basis Selection Haoran Wang, Chunfeng Yuan, Weiming Hu, Haibin Ling, Wankou Yang, and Changyin Sun.**

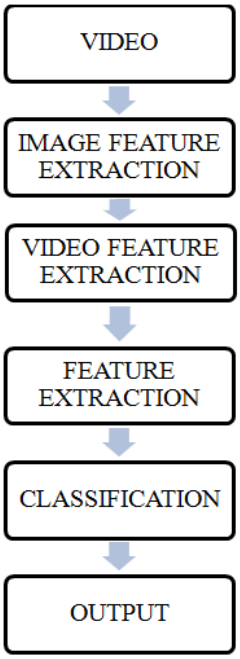
In this paper, we propose using high-level action units to represent human actions in videos and, based on such units, a novel sparse model is developed for human action recognition. There are three interconnected components in our approach. First, we propose a new context-aware spatialtemporal descriptor, named locally weighted word context, to improve the discriminability of the traditionally used local spatial-temporal descriptors. Second, from the statistics of the context-aware descriptors, we learn action units using the graph regularized nonnegative matrix factorization, which leads to a part-based representation and encodes the geometrical information. These units effectively bridge the semantic gap in action recognition. Third, we propose a sparse model based on a joint l2,1-norm to preserve the representative items and suppress noise in the action units. Intuitively, when learning the dictionary for action representation, the sparse model captures the fact that actions from the same class share similar units. The proposed approach is evaluated on several publicly available data sets. The experimental results and analysis clearly demonstrate the effectiveness of the proposed approach.

5. **Recognizing human actions in still images: a study of bag-of-features and part-based representations.**

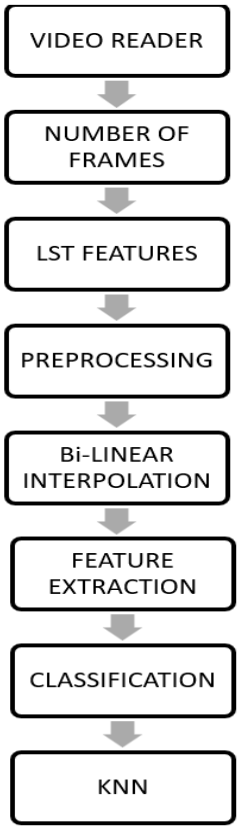
Recognition of human actions is usually addressed in the scope of video interpretation. Meanwhile, common human actions such as “reading a book”, “playing a guitar” or “writing notes” also provide a natural description for many still images. In addition, some actions in video such as “taking a photograph” are static by their nature and may require recognition methods based on static cues only. Motivated by the potential impact of recognizing actions in still images and the little attention this problem has received in computer vision so far, we address recognition of human actions in consumer photographs. We construct a new dataset with seven classes of actions in 968 Flickr images representing natural variations of human actions in terms of camera view-point, human pose, clothing, occlusions and scene background. We study action recognition in still images using the state-of-the-art bag-of-features methods as well as their combination with the part-based Latent SVM approach of Felzenszwalb et al. In particular, we investigate the role of background scene context and demonstrate that improved action recognition performance can be achieved by (i) combining the statistical and part-based representations, and (ii) integrating person-centric description with the background scene context. We show results on our newly collected dataset of seven common actions as well as

demonstrate improved performance over existing methods on the datasets.

BLOCK DIAGRAM



III. METHODS AND MATERIAL



1.NUMBER OF FRAME:

This formula above gives only an "estimate" of the number of frames. For fixed frame-rate files, the value will be within 1 of the actual number of frames due to rounding issues. Many video files are variable frame-rate and so for those files the actual number of frames can be more or less than the estimated value.

2. LST FEATURE(LAPLACIAN SMOOTHING TRANSFORM)

The Laplacian is a 2-D isotropic measure of the 2nd spatial derivative of an image. The Laplacian of an image highlights regions of rapid intensity change and is therefore often used for edge detection (see zero crossing edge detectors). The Laplacian is often applied to an image that has first been smoothed with something approximating a Gaussian smoothing filter in order to reduce its sensitivity to noise, and hence the two variants will be described together here. The operator normally takes a single gray level image as input and produces another gray level image as output. If a portion of the filtered, or gradient, image is added to the original image, then the result will be to make any edges in the original image much sharper and give them more contrast. This is commonly used as an enhancement technique in remote sensing applications.

3. PREPROCESSING

The purpose of the pre-processing stage is to remove unwanted effects such as noise from the image, and transform or adjust the image as necessary for further processing. The resolution of the image is reduced by a factor of four to 512•384 to speed up performance of the system. Also, the test images will be subjected to selective median filtering and unsharp masking to isolate noise which may have been accumulated during image acquisition and due to excessive staining.

4.BI-LINEAR INTERPOLATION

We can do this by looking at the values nearby. Methods include: – Nearest neighbor – just take the value of the closest neighbor – Bilinear – take a combination of the four closest neighbors – Bicubic – use the closest 16 neighbors (most computationally expensive, but best results).One of the simplest interpolation algorithms is nearest-neighbor interpolation. In this method, the fractional part of the pixel address is discarded, and the pixel brightness value at the resulting integral address in the source image is copied to the zoomed image. Because of the inexactness of the spatial correspondence between the two images, more copies will be made of certain pixels in the source image than of others. This can result in spatial distortion of features in the zoomed image, and nearest-neighbor interpolation is therefore unreliable for measurement purposes. For integral zoom factors that are even (such as 2X and 4X), nearest-neighbor interpolation produces the same results as the discrete replicating zoom algorithm.An interpolation technique that reduces the visual distortion caused by the fractional zoom calculation is the bilinear interpolation algorithm, where the fractional part of the pixel address is used to compute a weighted average of pixel brightness values over a small neighborhood of pixels in the source image. Bilinear interpolation produces pseudo resolution that gives a more aesthetically pleasing result, although this result is again not appropriate .

5.FEATURE EXTRACTION

In pattern recognition and in image processing, feature extraction is a special form of dimensionality reduction. When the input data to an algorithm is too large to be processed and it is suspected to be notoriously redundant, then the input data will be transformed into a reduced representation set of features. Transforming the input data into the set of features is called feature extraction. If the features extracted are carefully chosen it is expected that the features set will extract the relevant information from the input data in order to perform the desired task using this reduced representation instead of the full size input. Feature extraction involves simplifying the amount of resources required to describe a large set of data accurately. When performing analysis of complex data one of the major problems stems from the number of variables involved. Analysis with a large number of variables generally requires a large amount of memory and computation power or a classification algorithm which over fits the training sample and generalizes poorly to new samples. Feature extraction is a general term for methods of constructing combinations of the variables to get around these problems while still describing the data with sufficient accuracy.

MODULES:

1. Image Feature Extraction.
2. Video Feature Extraction.
3. Feature Extraction
4. Classifier.
5. KNN(K-nearest neighbor)

MODULE DESCRIPTION:

1. Image Feature Extraction:

In our method, we extract the image (static) feature from both images and key frames of videos. Considering computational efficiency, we extract key frames by a shot boundary detection algorithm[36]. The example of key frames 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.

2. Video Feature Extraction:

The video (motion) feature is extracted from the video domain and is combined with image

feature. Therefore, the image feature is a subset of the combined feature.

IV. RESULTS AND DISCUSSION

FEATURE EXTRACTION

In pattern recognition and in image processing, feature extraction is a special form of dimensionality reduction. When the input data to an algorithm is too large to be processed and it is suspected to be notoriously redundant, then the input data will be transformed into a reduced representation set of features. Transforming the input data into the set of features is called feature extraction. If the features extracted are carefully chosen it is expected that the features set will extract the relevant information from the input data in order to perform the desired task using this reduced representation instead of the full size input. Feature extraction involves simplifying the amount of resources required to describe a large set of data accurately. When performing analysis of complex data one of the major problems stems from the number of variables involved. Analysis with a large number of variables generally requires a large amount of memory and computation power or a classification algorithm which over fits the training sample and generalizes poorly to new samples. Feature extraction is a general term for methods of constructing combinations of the variables to get around these problems while still describing the data with sufficient accuracy.

1. Classifier:

The knowledge can be adapted based on such shared space of the common features, and then used to optimize the classifier A. In order to make use of unlabeled videos, a semi supervised classifier AB is trained based on the heterogeneous features in videos domain. We integrate the two classifiers into a joint optimization framework. The final recognition results of testing videos are improved by fusing the results of aforementioned two classifiers.

2. KNN Classification

The KNN binary (as two class) is given more accurate data classification which beneficial to select k as an odd number which avoids the irregular data. The KNN procedure is the technique in ML procedures: It is an object which classified through a mainstream selection of its neighbors, with the determination assigned occurrence for most mutual class amongst its k nearest neighbors (k is a positive integer, classically small). Classically Euclidean distance is used as the distance metric; however, this is only suitable for endless variables. In such situation as the classification of text, alternative metric, intersection metric or Hamming distance can be used. KNN is a new process that deliveries all available cases and categorizes novel cases built on an evaluation quantity (e.g., distance functions). KNN procedure is identical simple. It works built on a minimum distance from the interrogation instance to the training samples to regulate the K-nearest neighbors. The information for KNN procedure contains numerous attribute which will be used to categorize. The information of KNN can be any dimension scale from insignificant, to measurable scale.

V. CONCLUSION

The real-world applications which automatically label the beginning and ending of an action sequence. The system uses the proposed view-invariant features to address multiview action recognition from different perspectives for accurate and robust action recognition.

VI.FUTURE ENHANCEMENT

To emphasize that the reliability and computational efficiency of the proposed method allows the creation of an effective tool that can easily be incorporated in clinical practice.

VII. REFERENCES

[1]. B. Ma, L. Huang, J. Shen, and L. Shao, "Discriminative tracking using tensor pooling," IEEE Trans. Cybern., to be published,doi: 10.1109/TCYB.2015.2477879.

[2]. L. Liu, L. Shao, X. Li, and K. Lu, "Learning spatio-temporal representationsfor action recognition: A genetic programming approach," IEEETrans. Cybern., vol. 46, no. 1, pp. 158–170, Jan. 2016.

[3]. A. Khan, D. Windridge, and J. Kittler, "Multilevel Chinese takeawayprocess and label-based processes for rule induction in the contextof automated sports video annotation," IEEE Trans. Cybern., vol. 44,no. 10, pp. 1910–1923, Oct. 2014.

[4]. H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluationof local spatio-temporal features for action recognition," in Proc. Brit.Mach. Vis. Conf., London, U.K., 2009, pp. 124.1–124.11.

[5]. L. Shao, X. Zhen, D. Tao, and X. Li, "Spatio-temporal Laplacianpyramid coding for action recognition," IEEE Trans. Cybern., vol. 44,no. 6, pp. 817–827, Jun. 2014.

[6]. M.-Y. Chen and A. Hauptmann, "MoSIFT: Recognizing human actionsin surveillance videos," School Comput. Sci., Carnegie Mellon Univ.,Pittsburgh, PA, USA, Tech. Rep. CMU-CS-09-161, 2009.

[7]. M. Yu, L. Liu, and L. Shao, "Structure-preserving binary representationsfor RGB-D action recognition," IEEE Trans. Pattern Anal. Mach. Intell.,to be published, doi: 10.1109/TPAMI.2015.2491925.

[8]. L. Shao, L. Liu, and M. Yu, "Kernelized multiview projection for robust action recognition," Int. J. Comput. Vis., 2015,doi: 10.1007/s11263-015-0861-6.

[9]. C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vectormachines," ACM Trans. Intell. Syst. Technol., vol. 2, no. 3, pp. 1–27,Apr. 2011.

[10]. Y. Han et al., "Semisupervised feature selection via spline regressionfor video semantic recognition," IEEE Trans. Neural Netw. Learn. Syst.,vol. 26, no. 2, pp. 252–264, Feb. 2015

Cite this article as :

Paruchuri Yogesh, R. Dillibabu, Palaniappan P, Prof. L. Ashok Kumar, Prof. Navarajan, " Robust Feature Based Automated Multi View Human Action Recognition System Using Machine Learning, International Journal of Scientific Research in Science, Engineering and Technology(IJSRSET), Print ISSN : 2395-1990, Online ISSN : 2394-4099, Volume 6, Issue 2, pp.52-58, March-April-2019. Available at doi : https://doi.org/10.32628/IJSRSET1961166

Journal URL : http://ijsrset.com/IJSRSET1961166