A Review on Human Action Recognition

[1]Shraddha Marotkar, [2]Prof. A. B. Kharate

[1]PG Student, [2]Professor
Department of Electronics and Telecommunication Engineering,
Amravati University, Amaravati, India

Abstract: Human-to-human interaction and interpersonal relations Human activity recognition plays a significant role. Because to identity of a person, their personality, and psychological state it provides information, it is difficult to extract. Another person's activities is one of the main subjects of study of the scientific areas of computer vision the human ability to recognize and soft computing learning. Robotics for human behavior characterization, require a multiple activity recognition system, many applications, including video surveillance systems, human computer interaction, and. human activity recognition is an important research direction In video analysis, a large number of papers have been published on human activity recognition in video and image sequences In the past, including methods, systems, and quantitative evaluation of the performance of human activity recognition In this paper, we provide a comprehensive survey of the recent development of the techniques, for the video improve classification, interpretation, and performance The experimental results show that our method can significantly.

Index Terms: Human Action Recognition; Feature Extraction; Feature Reduction; Classifiers; KTH Dataset.

I. INTRODUCTION

In digital video content acquisition and visualization makes it increasingly challenging to find, organize, and access visual information in the continual rapid growth. Content-based image retrieval (CBIR), has progressed over many decades Research to better represent and understand visual content. Global and local features are CBIR is based on two types of visual features: In visual content as a whole Global feature based algorithms aim at recognizing concepts, they are often not directly related to any high-level semantics. Local features are an alternative choice and have several advantages over global features is the main drawback is that. Local feature contain the rich local information in an image Local feature algorithms focus mainly on key points, the salient image patches. Using various detectors, e.g., Harris corner [2] and difference of Gaussian (DoG) [3] these can be automatically detected. Low-level visual descriptor for the scale invariant feature transform (SIFT) [4] is a promising, translation, and rotation, and as well as partially invariant to illumination changes and affine projections, low-level visual descriptor is invariant to scaling. For an image, several challenges need to be overcome To obtain the high-level semantics. First, visual data need to be analyzed and transformed into a format that represents the visual content effectively, the camera is stationary, assumes that.

Problem Statement

Identifying human activities from Video dataset has proved to be complex task due to large dimensions of dataset. Various Machine learning techniques have been used previously to identify human activities. We have proposed a system that reduces dimensions of Smartphone dataset and uses Machine learning algorithms in an optimized manner to produce efficient result.

Related Work

Xian-gan Chen [IEEE Int. Conf. On Electric Information and Control Engineering, Wuhan, 2011] performed automatic recognition of human actions using Support Vector Machine (SVM). The focus is on simple actions in simple background. The morphological gradient of an input video is performed to get silhouette of the human body. The second step is to perform pooling operation on the output obtained from the morphological gradient step. This step helps increasing the invariance of the shape. In the third step, the sequence is fed to Pyramid Histogram of Oriented Gradient (PHOG). In this step, canny edge detector is used to extract the edges of the video frames. Finally SVM is used to classify different actions. The results are very well compared to other approaches following SVM. Local learning method is used in [8] to automate human action recognition in video clips. The local learning method reduces the problem of the variations in an action class. To speed it up, local classifiers are trained jointly rather than iteratively. The main steps are feature extraction, mapping features to code words and boosting classifier for action recognition.

Optical flow and silhouette-based features were used for view-invariant action recognition in [M. Ahmad and S.-W. Lee 2006. ICPR 2006. 18th International Conference], and principal component analysis (PCA) was used for reducing the dimensionality of the data. , coarse silhouette features, radial grid-based features and motion features were used for multi view action recognition. Another method for viewpoint changes and occlusion-handling was proposed. This method used the histogram of oriented gradients (HOG) features with local partitioning, and obtained the final results by fusing the results of the local classifiers. A novel motion descriptor based on motion direction and histogram of motion intensity was proposed for multi view action recognition followed by a support vector machine used as a classifier. Another method based on 2D motion templates, motion history images, and histogram of oriented gradients was proposed i. A hybrid model which combines convolution neural networks (CNN) with hidden Markov model (HMM) was used for action classification. In this method, the CNN was used to learn the effective and robust features directly from the raw data, and HMM was used to learn the statistical dependencies over the contiguous actions and conclude the action sequences.

A key strand of related work is video feature representations for machine learning. Local video representations have been successfully applied to for human action recognition applications [Laptev et al., 2004, Laptev et al.,2008]. There are typically three steps to extract representations - interest points detection, descriptors extraction and aggregation. Interest points denote points in an image or video which are likely to give strong signal information, such as a corner of junction. Interest point detectors select space-

time locations and scales from videos and distinguish the sparseness of the interest points. Then, the local descriptors encode appearances and motions around the selected interest points based on the measurement of space-time gradients and optical flow. For interest points detection, space-time interest point (STIP) was proposed by [Laptev, 2005], which is an extension of Harris corner detection [Harris and Stephens, 1988] by finding out significant variations in both spatial and temporal domains. Another commonly used detector is the Hessian 3D detector, which is also extended from 2D counterparts [Willems et al., 2008]. In relation to STIP and Hessian3D detectors, the cuboid detector was proposed to overcome the limitation of STIP that insufficient interest points may be detected to represent actions [Dollár et al., 2005]. To obtain the local descriptors from detected interest points, HOF was presented to encode pixels-level motions from optical flow fields through local patches as descriptors [Laptev et al., 2008]. HOG3D was proposed to describe motion features in human actions [Klaser et al., 2008], which is an extension of HOG used in human detection from images [Dalal and Triggs, 2005]. Another robust descriptor is motion boundary histogram (MBH) that compute gradients over optical flow fields. Notably, trajectory based methods such as iDT track interest points to form trajectory-based 3D volumes, computing HOG/HOF and MBH as descriptors for each trajectory [Wang and Schmid, 2013]. In some supervised machine learning classifiers such as SVM, the size of the input vector should be fixed-length vectors. However, the number of local descriptors varies for each video. Therefore, BoVWs are used to aggregate the sets of local descriptors into fixed length vectors, which can be fed into supervised classifiers [Laptev et al., 2004, Dollár et al., 2005, Laptev et al., 2008, Klaser et al., 2008]. Generally, in BoVWs, each video is represented as a histogram of the frequency of the visual words appearing as the closest match to the local features in the codebook. In traditional transfer learning approaches, a discriminative transfer learning method based on least squares SVM (LS-SVM) was proposed by [Tommasi et al., 2010], to learn object categories from few samples. Domain transfer SVM (DT-SVM) was introduced by [Duan et al., 2009] to minimise the data distribution mismatch between the target and source domain using the maximum mean discrepancy (MMD) criterion that measures the distribution similarity of two domains. Additionally, the transfer learning adaptive boosting (TrAdaBoost) is introduced by [Dai Wenyuan et al., 2007], that is the extension of adaptive boosting (AdaBoost) [Freund and Schapire, 1997], to reduce the weighted training error on the data with different distribution, and in the meantime preserving the properties of AdaBoost. Furthermore, the A-SVM transfer learning variant of SVM was originally introduced by [Yang et al., 2007]. It learns from the source model by regularising the distance between the learned model and source model, used in video concept detection application. The work of [Aytar and Zisserman, 2011] then focused on object detection, referencing the use of A-SVM, and proposing PMT-SVM, which can increase the amount of transfer without penalising margin maximisation. In this work, we will first examine two video representations that are STIP+HOG/HOF and iDT, used in conjunction with two classifier-based transfer learning approaches that are A-SVM and PMT-SVM. Both transfer learning approaches are presented for binary classification tasks only. We will compare their performance for the tasks of distinguishing multiple classes of human actions.

II. STATE OF THE ART

For object recognition the advantages of local features versus global features extraction and visual content categorization have been summarized by Lee [8]. Lowe [3]features extraction is invariant to image scaling, translation, rotation, and partially invariant to illumination changes and affine projections proposed a technique for visual features extraction . This technique is called scale invariant feature transform (SIFT). Ke et al. [9] improved the SIFT technique by applying principal components analysis (PCA) to reduce the dimensions of SIFT descriptors in video. local descriptors are more distinctive, more robust to image deformations, and more compact, compared to the standard SIFT representation, increasing image retrieval accuracy and comparing speed. This paper presents a set of techniques integrated into a low-cost PC based real time visual surveillance system for simultaneously human motion detection, tracking people, and analysis their activities in monochromatic video. video tracking in a single camera is performed using background subtraction, followed by region correspondence.

Proposed System:

The most powerful image processing system is the human brain together with the eye, but here we tried to develop this system keeping this fact in mind. The system receives a particular video, enhances, divide them into each different frames and stores those derived images at enormous rates of speed in a particular folder for later usage. The user will send videos and according to the specification they will be modified, and converted top frames and then further detections will be performed as per the user's choice viz. to detect human action or to use the system for detecting the speed of particular objects in the frame, or to merge both the concepts and use as an encapsulated software experience.

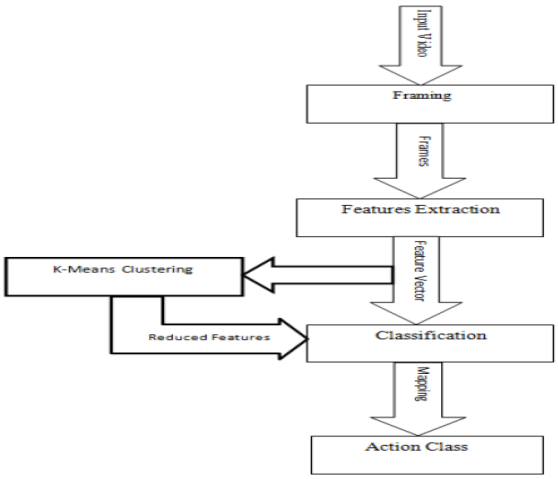
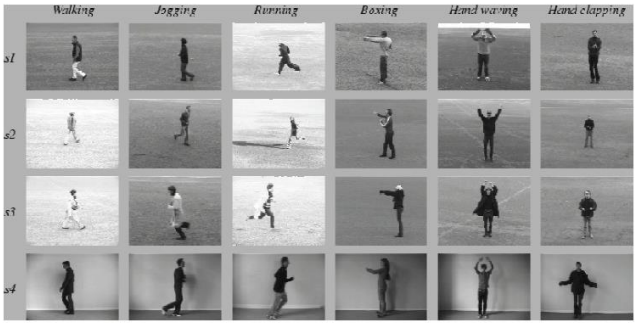


Fig. 1 Flow Diagram of Proposed Approach

Dataset

To evaluate the performance of our approach, we use KTH dataset [11]. It contains six types of human actions including walking, jogging, running, boxing, hand waving and hand clapping, performed several times by 25 subjects in four different scenarios: outdoors, outdoors with scale variation, outdoors with different clothes, and indoors. Currently the database contains 2391 sequences. All sequences were taken over homogeneous backgrounds with a static camera with 25fps frame rate. The sequences were down sampled to the spatial resolution of 160x120 pixels and have a length of four seconds in average.



KTH Sample Dataset

Modules:

Action Recognition: from video data the goal of the action recognition is an automated analysis of ongoing events. This module recognizes the action based on the objects, rather human model and correctly recognizes some basic human action like,

- Walking
- Running
- Surfing
- Clapping
- Boxing
- Waving
- Jogging
- Cycling

System Architecture:

Behavior, and more views of a system. Organized in a way supports reasoning about the structures and behaviors of the system An architecture description is a formal description and representation of a system.

III. REPRESENTATION FRAMEWORK

After inputting an original video, image segmentation techniques are employed to theoretically extract all the pixels belonging to object. Without knowing the action of the object, the aim of the segmentation step here is potentially contain object to initially divide the pixels of each frame from a video clip into two classes of regions which do not contain text and regions which each video frame can be considered to consist of homogeneous segments After segmentation. By filtering the monochrome segments whose widths and heights are too large or too small to be instances of video characters, binarization and dilation are performed. Finally, image enhancement techniques such as contrast analysis and aspect ratio restriction can be employed to obtain better preprocessing results for further video object detection or recognition. the manually added graphics texts usually have a strong contrast between the character regions and their surrounding backgrounds for highlights and human reading in example. This property makes contrast analysis helpful to decide whether directly discarding a video segment or instead sending the segment for further video object detection and character recognition. This chapter gives a brief overview of the above mentioned preprocessing techniques that are often used in video object detection.

IV. CONCLUSIONS

This paper gives a brief overview of preprocessing techniques related to human action recognition. We introduce image pre-processing operators and color based and texture-based preprocessing techniques which are most frequently used in video analysis. Since image segmentation potentially helps reduce the searching space and more important, provides contextual hints in complex video scenes for human action recognition, we then give a brief introduction on automatic segmentation technique. we introduce how to compute motions from video frames and how video speed detection benefits from motion analysis.

REFERENCES

1. "Improved SIFT-features matching for object recognition," F. Alhwarin, C. Wang, D. Ristic-Durrant, and A. Graser, in *Proc. Int. AcademicConf. Vision of Computer Science-BSC*, 2008, pp. 179–190.
2. "A combined corner and edge detector", C. Harris and M. Stephens, in *Proc. 4th Alvey Vision Conf.*, 1988, pp. 147–151.
3. "Object recognition from local scale-invariant features", D. G. Lowe, in *Proc. Int. Conf. Computer Vision*, 1999, vol. 2, pp.
4. "Distinctive image features from scale-invariant keypoints", D. G. Lowe, *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
5. " Recognizing Human Actions: A Local SVM Approach"**,** Christian Sch¨uldt, Ivan Laptev ,Barbara Caputo Computational Vision and Active Perception Laboratory (CVAP)
6. "Real Time Video Surveillance of Human Activity" Larry Davis,Computer Vision Laboratory, University of Maryland, College Park, Maryland
7. "Retrieving actions in movies**",** Ivan Laptev and Patrick P´erez , RISA / INRIA Rennes, Campus universitaire de Beaulieu
8. **"**Local and global feature extraction for face recognition", Y. Lee, K. Lee, and S. Pan, in *Proc. 5th Int. Conf. Audio- and Video-Based Biometric Person Authentication*, 2005, pp. 219–228.

9. "PCA-SIFT: A more distinctive representation for local image descriptors", Y. Ke and R. Sukthankar, in *Proc. 2004 IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, 2004, vol. 2, pp.513–506.
10. "DETECTING HUMAN ACTION IN ACTIVE VIDEO**",** *Hao Jiang, Ze-Nian Li and Mark S. Drew,*School of Computing Science, Simon Fraser University, Vancouver, BC, Canada V5A 1S6
11. "Motion Detection, Tracking and Classification for Automated Video Surveillance**",** Neha Gabal, Neelam Barak and Shipra Aggarwal,Department of Mechanical and Automation Engineering, IGDTUW, New Delhi, India
12. "AUTO MOTOR VEHICLE APPLICATIONS APPLIED MATERIAL HANDLING IN MATLAB", Performed by: Lê TiếnSĩ. Faculty of Electrical and Electronics Engineering, University of Technical Education TP.Asst. Researcher @ Ho Chi Minh City University of Technology and Education, VN
13. "HUMAN ACTION DETECTION USING KTH DATASET"**,** Schüldt, Christian, Ivan Laptev, and Barbara Caputo. "Recognizing human actions: a local SVM approach." Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on. Vol. 3. IEEE, 2004.
14. X. Chen, J. Liu, Z. Gao, and H. Liu, "Recognizing human actions from video sequences using invariant shape", IEEE Int. Conf. On Electric Information and Control Engineering, Wuhan, pp. 1564–1567, April 2011.
15. M. Ahmad and S.-W. Lee, "Hmm-based human action recognition using multiview image sequences," in Pattern Recognition, 2006. ICPR 2006. 18th International Conference on, vol. 1, pp. 263–266, IEEE.