Human Action Recognition in Video Forms

Anju Marium Abraham[1] Smita C Thomas[2]
[1,2]Mount Zion College of Engineering Kadamanitta, Pathanamthitta, India

Abstract— From various scattered similar the action is recognized on the basis of space time and volume in video forms, which can be fully characterized by a linear rank decomposition. Recurrence plot theory is applied, and introducing the concept of Joint Self-Similarity Volume (Joint-SSV) to model this scattered action in various forms, and hence rank-1tensor approximation of the Joint-SSV is applied to obtain very low-dimensional descriptors that very correctly characterize an action in a video sequence id used [5].The descriptor vectors make it more possible to recognize actions without explicitly aligning the videos in time in order to compensate for speed of execution or differences in video frame rates. The method is generic, in the sense that it can be applied using different level features, such as tracked points, 2D block diagrams, histogram of oriented gradients. Therefore, the method does not necessarily require explicit tracking of features in the space-time volume.

Key words: Recognition, Video Forms

I. INTRODUCTION

Action recognition has continued to be an active area of research and has thus rightfully attracted much attention from the researchers over the years. Important application domains, such as automatic video indexing and archiving, video surveillance, human-computer interaction, augmented reality, user interface design, and human factors would benefit immensely from a robust and efficient solution to this problem. There are many factors that make this a challenging problem, including the large variations in performing an action by different people, whether by varying the postures, or the execution speed, illumination variations in the sequences, occlusions and disocclusions, distracting background motions, and perspective effects and camera motion. As a consequence, current methods often resort to restricted and simplified scenarios with simple backgrounds, simpler kinematic action classes, static cameras or limited view variations. Various approaches have been proposed over the years for action recognition [1]. On the basis of representation, they can be categorized as: time evolution of human silhouettes, space-time shapes, dense trajectories, and local 3D patch analysis, generally coupled with some machine learning techniques. All these works rely primarily on effective feature extraction. These feature extraction methods can be roughly divided into the following four categories: motion based, appearance based, space-time volume based, and space-time interest points or local features based. Motion based methods generally compute optical flow from a given action sequence, followed by appropriate feature extraction. However, these methods are known to be very susceptible to noise and easily lead to inaccuracies. Appearance based methods are prone to differences in appearance between the training dataset and the testing sequences. Volume or shape based methods mostly require highly detailed silhouette extraction, which may not be possible in real-world noisy video dataset. In comparison with these approaches, the space-time interest point (STIP) based methods are more robust to noise and camera movement and also seem to work quite well with low resolution inputs. However, these methods rely solely on the discriminative power of individual local space-time descriptors [4]. Information related to the global spatio-temporal distribution is ignored. Thus due to lack of this temporal information, smooth motions cannot be captured using STIP methods. In addition, issues like optimal space-time descriptor selection and codebook clustering algorithm selection have to be addressed, with fine-tuning various parameters, which is highly data dependent. The notion of "self-similarity" has received significant attention, recently. The work in describes a gait recognition technique based on the image self-similarity of a walking person and classify the movement patterns of different people. Some works, also show the effective use of the self-similarity in recognizing different types of biological periodic motions. The authors exploit the notion of image self-similarities, as proposed by. For a given action sequence, first extracts some low level features. The distances between extracted features for all pairs of time frames are computed and this results in a Self-Similarity Matrix (SSM). Each action sequence is thus reduced to a 2D SSM, and the authors then proceed to extracting some useful features from these SSMs and use it to train the action recognition system.

The concept of self-similarity is also closely related to the statistical co-occurrence of pixel intensities across images captured by Mutual Information. The paper in proposed an image matching method based on internal self-similarity property of images, and explored various definitions of self-similarity to find the best one for image matching. In terms of video analysis, there are mainly two types of self-similarity based descriptors in the literature, namely the local self-similarity descriptor (LSS) and the global self-similarity descriptor (GSS). In [26] the authors introduced the concept of LSS descriptors that capture internal geometric layouts of local self-similarities with videos, and these descriptors are estimated on a dense grid of points in the video data at multiple scales. This type of descriptor captures the internal layout of local regions and can be compared across images which appear substantially different at the pixel level. Built upon this method, explored instead the structure of similarities between all pairs of time-frames in a sequence, and made only mild assumptions about the rough localization of a person in the frame, instead of relying on structure recovery and correspondence estimation. Myriads of other applications are also proposed based on this LSS descriptor. In contrast, in the authors explored instead the global self-similarity and its advantages over the local ones, and proposed two types of global descriptors: the bag-of-correlation-surfaces and self-similarity hypercube.

II. LITERATURE SURVEY

In this paper [1], With the development of advanced security systems, human action recognition in video sequences has become an important research topic in computer vision,

whose aim is to make machines recognize human actions using different types of information, especially the motion information, in the video sequences. The basic process for this problem can be divided into three issues: First, how to detect the existence of human actions? Second, how to represent human actions? Lastly, how to recognize these actions? Mainly focus on the second issue, that is, how to represent human actions after having detected their existence. In our approach, we model each video sequence as a collection of so-called motion images (MIs), and to model the action in eachMI, we propose a novel motion-based representation called motion context (MC), which is insensitive to the scale and direction of an action, to capture the distribution of the motion words (MWs) over relative locations in a local region around a reference point and thus summarize the local motion information in a rich, local 3D MC descriptor.

In this paper [2], nowadays, there is an ever-increasing migration of people to urban areas. Health care service is one of the most challenging aspects that is greatly affected by the vast in_ux of people to city centres. Consequently, cities around the world are investing heavily in digital transformation in an effort to provide healthier ecosystems for people. In such a transformation, millions of homes are being equipped with smart devices (e.g., smart meters, sensors, and so on), which generate massive volumes of _fine-grained and indexical data that can be analysed to support smart city services ,propose a model that utilizes smart home big data as a means of learning and discovering human activity patterns for health care applications, the use of frequent pattern mining, cluster analysis, and prediction to measure and analyse energy usage changes sparked by occupants' behaviour. Since people's habits are mostly identified by everyday routines, discovering these routines allows us to recognize anomalous activities that may indicate people's difficulties in taking care for themselves, such as not preparing food or not using a shower/bath, addresses the need to analyze temporal energy consumption patterns at the appliance level, which is directly related to human activities.

In this paper[3], Within the field of action recognition, features and descriptors are often engineered to be sparse and invariant to transformation. While sparsity makes the problem tractable, it is not necessarily optimal in terms of class separability and classification. This paper proposes a novel approach that uses very dense corner features that are spatially and temporally grouped in a hierarchical process to produce an over complete compound feature set. Frequently reoccurring patterns of features are then found through data mining, designed for use with large data sets. The novel use of the hierarchical classifier allows real time operation while the approach is demonstrated to handle camera motion, scale, human appearance variations, occlusions and background clutter. use data mining within the action recognition field to allow a multi stage classifier to be learnt from a large set of simple features. Initially, a very localised neighbourhood grouping is used to form a compound grouping of features. The neighbourhood is hierarchically increased until the final stage uses only the relative position of groupings to provide scale invariance. The novel hierarchical approach allows for real time operation and out performs other state of the art approaches on the KTH dataset. Finally we present results on two more challenging sequences. Multi-KTH to demonstrate its performance at classifying and localising multiple actions in noisy cluttered scenes containing camera motion and the extremely challenging Hollywood dataset of which contains segments extracted from a variety of movies.

III. Conclusion

The recurrence plot theory to define a tensor representation of the dynamics of an action in video data, which we refer to as a Joint Self-Similarity Volume, the Joint-SSV is sparse when applied to videos of human actions. In other words, it can be for the most part characterized by its rank-1 subspace representation. Therefore, by exploiting this sparseness, we reduce the high-dimensional recognition problem to a linear low-dimensional matching problem in a rank-1 subspace, without compromising our recognition accuracy. A particular feature of our approach is that it leads to a generic solution to this problem in the sense that our solution is independent of the type of input features, i.e. tracked points in a motion capture dataset, manually marked points, automatically extracted silhouettes, Histogram of Gradient (HoG) feature vectors, optical flow, etc. For reducing the dimensionality of the Joint-SSV, introduced a new rank-1 tensor approximation algorithm that relies on an alternating least squares approach to find the optimal rank-1 decomposition.

Acknowledgment

We would like to thank, first and foremost, Almighty God, without his support this work would not have been possible. We would also like to thank all the faculty members of Mount Zion College of engineering, for their immense support.

References

[1] Z. Zhang, Y. Hu, S. Chan, and L.-T. Chia, "Motion context: A new representation for human action recognition," in Proc. 10th Eur. Conf. Comput. Vis. (ECCV), 2008, pp. 817–829.
[2] A. A. Efros, A. C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in Proc. 9th IEEE ICCV, Oct. 2003, pp. 726–733.
[3] A. Gilbert, J. Illingworth, and R. Bowden, "Fast realistic multi-action recognition using mined dense spatio-temporal features," in Proc. IEEE 12th Int. Conf. Comput. Vis., Sep. /Oct. 2010, pp. 925–931.
[4] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in Proc. Brit. Mach. Vis. Conf. (BMVC), 2009, pp. 124.1–124.11.
[5] M. F. Abdelkader, W. Abd-Almageed, A. Srivastava, and R. Chellappa, "Silhouette-based gesture and action recognition via modelling trajectories on Riemannian shape manifolds," Comput. Vis. Image Understand., vol. 115, no. 3, pp. 439–455, 2011.

