ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

IJESRT
INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

REVIEW PAPER ON REAL TIME HUMAN ACTION RECOGNITION USING SHAPE FEATURES HISTOGRAMS

Vipin Kumar Batra[*1] & Mrs. Priyanka Gaur[2]
[*1]M. Tech Scholar, E.C.E., Advanced Institute of Technology, and Management, AITM, Palwal
[2]Assistant Professor, E.C.E., Advanced Institute of Technology, and Management, AITM, Palwal

DOI: 10.5281/zenodo.1247295

ABSTRACT

Human Activity Recognition (HAR) is the understanding of human behaviour from data captured by pervasive sensors, such as cameras or wearable devices. It is a powerful tool in medical application areas, where consistent and continuous patient monitoring can be insightful. Wearable devices provide an unobtrusive platform for such monitoring, and due to their increasing market penetration, feel intrinsic to the user. This daily integration into a user's life is crucial for increasing the understanding of overall human health and wellbeing. This is referred to as the "quantified self" movement. Wearables, such as actigraph accelerometers, generate a continuous time series of a person's daily physical exertion and rest. This ubiquitous monitoring presents substantial amounts of data, which can (i) provide new insights by enriching the feature set in health studies, and (ii) enhance the personalization and effectiveness of health, wellness, and fitness applications. By decomposing an accelerometer's time series into distinctive activity modes or actions, a comprehensive understanding of an individual's daily physical activity can be inferred. The advantages of longitudinal data are however complemented by the potential of noise in data collection from an uncontrolled environment. Therefore, the data sensitivity calls for robust automated evaluation procedures.

I. INTRODUCTION

Human activity recognition plays a significant role in human-to-human interaction and interpersonal relations. Because it provides information about the identity of a person, their personality, and psychological state, it is difficult to extract. The human ability to recognize another person's activities is one of the main subjects of study of the scientific areas of computer vision and machine learning. As a result of this research, many applications, including video surveillance systems, human-computer interaction, and robotics for human behavior characterization, require a multiple activity recognition system. Among various classification techniques two main questions arise: "What action?" (i.e., the recognition problem) and "Where in the video?" (i.e., the localization problem). When attempting to recognize human activities, one must determine the kinetic states of a person, so that the computer can efficiently recognize this activity. Human activities, such as "walking" and "running," arise very naturally in daily life and are relatively easy to recognize. On the other hand, more complex activities, such as "peeling an apple," are more difficult to identify. Complex activities may be decomposed into other simpler activities, which are generally easier to recognize. Usually, the detection of objects in a scene may help to better understand human activities as it may provide useful information about the ongoing event (Gupta and Davis, 2007).

Most of the work in human activity recognition assumes a figure-centric scene of uncluttered background, where the actor is free to perform an activity. The development of a fully auto- mated human activity recognition system, capable of classifying a person's activities with low error, is a challenging task due to problems, such as background clutter, partial occlusion, changes in scale, viewpoint, lighting and appearance, and frame resolution. In addition, annotating behavioral roles is time consuming and requires knowledge of the specific event. Moreover, intra- and interclass similarities make the problem amply challenging. That is, actions within the same class may be expressed by different peo- ple with different body movements, and actions between different classes may be difficult to distinguish as they may be represented by similar information. The way that humans perform an activity depends on their habits, and this makes the problem of identifying the underlying activity quite difficult to determine. Also, the construction of a visual model for learning and analyzing human movements in real time with inadequate benchmark datasets for evaluation is challenging tasks.

II. RELETED WORK

This section presents state-of-the-art methods for multiview action recognition based on a 2D approach. These methods extract features from 2D image frames of all available views and combine these features for action recognition. Then, classifier is trained using all these viewpoints.

After training the classifier, some methods use all viewpoints for classification [10], while others use a single viewpoint for classification of a query action [12]. In both cases, the query view is part of the training data. However, if the query view is different than the learned views, this is known as cross-view action recognition. This is even more challenging than the multiview action recognition [9].

Different types of features—such as motion features, shape features, or combination of motionand shape-based features—have been used for multiview action recognition. In [8], silhouette-based features were acquired from five synchronized and calibrated cameras. The action recognition from multiple views was performed by computing the R transform of the silhouette surfaces and manifold learning. In [2], contour points of the human silhouette were used for pose representation, and multiview action recognition was achieved by the arrangements of multiview key poses. Another silhouette-based method was proposed in [13] for action recognition from multiple views; this method used contour points of the silhouette and radial scheme for pose representation. Then, model fusion of multiple camera streams was used to build the bag of key poses, which worked as a dictionary for known poses and helped to convert training sequences into key poses for a sequence-matching algorithm. In [13], a view-invariant recognition method was proposed, which extracted the uniform rotation-invariant local binary patterns (LBP) and contour-based pose features from the silhouette.

The classification was performed using a multiclass support vector machine. In [4], scale-invariant features were extracted from the silhouette and clustered to build the key poses. Finally, classification was done using a weighted voting scheme.

An optical flow and silhouette-based features were used for view-invariant action recognition in [6], and principal component analysis (PCA) was used for reducing the dimensionality of the data. In [3], coarse silhouette features, radial grid-based features and motion features were used for multiview action recognition. Another method for viewpoint changes and occlusion-handling was proposed in [2]. This method used 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 in [7] for multiview 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 in [8]. A hybrid CNN–HMM model which combines convolution neural networks (CNN) with hidden Markov model (HMM) was used for action classification [7]. 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 subactions and conclude the action sequences.

Piles of work have been done in human action recognition and localization to save manual work and accelerate processing efficiency. Although there is no perfect algorithm to deal with all the problems occurring in action recognition, such as viewpoint change, complex background activities and partial occlusion, yet the current work has shown quite promising results under different scenarios. One branch of work is to utilize the global information of the subject subtracted from video data, for example the correlating optical flow measurements from low resolution videos proposed by *Efros et al.* [19] which segment and stabilize each human figure and annotate each action in the resulted spatial temporal volume. Another part of the work attempts to track the body parts and use the motion trajectories to discriminate different actions [20, 21]. In their implementations, certain feature points are located in a frame-by-frame manner, and the tracks of these points show many discriminative properties, such as position, velocities and appearance. However, the methods mentioned above are sensitive to partial occlusion and use much redundant information that is computationally expensive. The drawbacks of these methods accelerate the prosperity of the part-based approaches, especially the space-time interest point detectors proposed by *Laptev et al.* [5] and *Dollár et al.* [6].

Laptev et al. [5] propose a space-time interest point detector based on the idea of the *Harris* and *Förstner* interest point operators. Their approach detects local structures in space-time where the image values have significant local variations in both space and time. *Dollár et al.* [6] use a set of separable linear filters detecting interest points with strong motion. This method is designed to respond to complex motion of local regions and the space-time corners. It detects more number of interest points than *Laptev*'s approach, which makes it more reliable in processing videos with limited frames. *Ke et al.* [7] apply spatial temporal volumetric features that efficiently scan video sequences in space and time and detect interest points over the motion vectors.

The detected spatial temporal interest points can be used to construct 3D shape context descriptors based on the concept that one certain action provides similar structural distribution of interest points [38]. This structural information refers to the configuration of entire shape with regard to a reference point [35]. On the other hand, *Dollár et al.* [6] extract spatial temporal video patches (cuboids) around each interest point, which contain the appearance information of video event, and can be further processed to provide unique features. This appearance information can be brightness gradient, optical flow and graphical shape etc.

However, we discovered that more compact and informative description approaches can be applied in the human action recognition area. Therefore, in the following sections the main objective is to present and validate the projected 3DSC in action recognition problem; and also to introduce transform domain techniques into this field, due to their great performance in signal processing; and finally to prove correlogram outperforms histogram in representing features.

III. SHAPE CONTEXT

The definition behind shape context is mainly referring to two aspects as elaborated in [30] and [31]. The first aspect is feature based approach, which takes advantage of the spatial arrangements of the extracted features, for instance the silhouette elements or junctions. The second aspect is brightness based approach which uses pixel values directly. Many years of work have been done based on the first method mentioned above. *Sharvit et al.* [32] attempt to capture the part structure of the shape in the graph structure of the skeleton. *Gdalyahu et al.* [33] use 1D silhouette curves for matching between two objects. However, silhouette based methods have intrinsic drawbacks due to the fact that silhouette ignores the internal contour and is very hard to extract from complex background. One alternative is to consider the object shape as a set of points and use edge detector to extract points as explained in [34]. *Belongie et al.* [35] propose to capture a subset of edge points and for each point build a shape context which is a histogram of the relative coordinates of the remaining points in shape matching and object recognition. Derived from the idea of [35], *Kortgen et al.* [36] extend 2D Shape Context into 3 dimensional object matching. *Shao and Du* [37] further extend *Belongie et al.*'s idea into 3D video sequences, which is to build descriptors containing the spatial temporal distribution of the remaining interest points regarding to a reference point. This approach, named as 3 Dimensional Shape Context (3DSC), performs well in action recognition with scale and translation invariance.

IV. TRANSFORM BASED DESCRIPTORS

Transform domain techniques have been widely used in the image processing field, such as image compression, enhancement and segmentation etc. There are three well known transforms in this area, which are Discrete Fourier Transform (DFT), Discrete Cosine Transform (DCT) and Discrete Wavelet Transform (DWT). Fourier Transform is to analyze a signal in the time domain for its frequency content and translate it into a function in the frequency domain. The DFT estimates the Fourier transform of a function from a finite number of its sampled points [24]. DCT is similar to DFT, but only using real data with even symmetry. DWT decomposes a discrete signal into a set of discrete basis functions (wavelets) instead of sine and cosine waves used in Fourier Transform, and it captures both the frequency information and the space information.

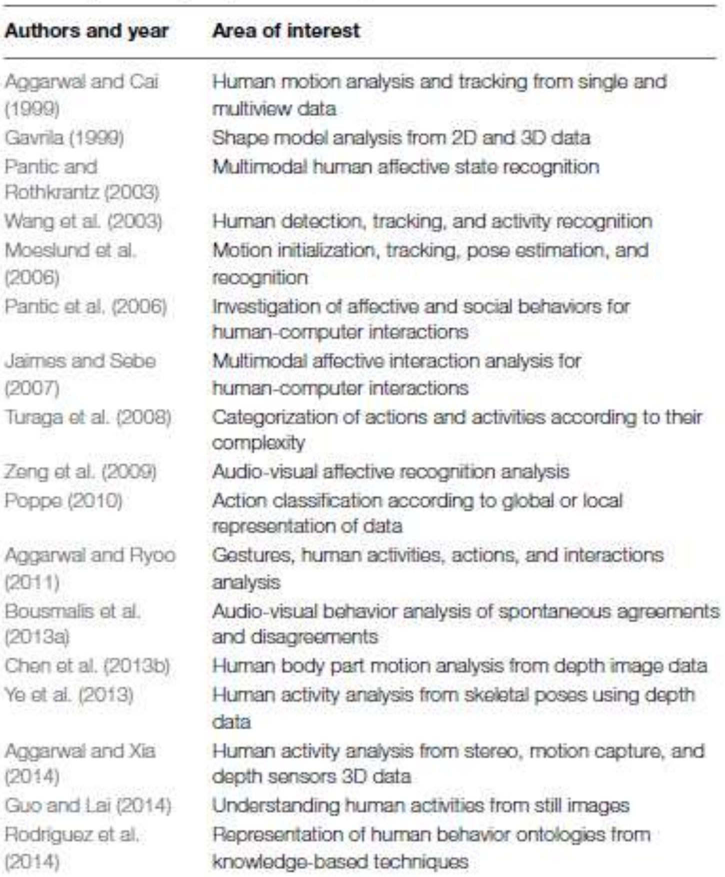
Recently, DCT and DWT have been widely exploited by modern image and video coding standards, such as JPEG, MPEG etc. [9, 15]. *Smith et al.* [50] propose a texture classification method based on the variance and the mean absolute values of the DCT coefficients calculated over the entire image. *Climer et al.* [51] propose a quadtreestructure- based method using the DCT coefficients on the nodes of the quadtree as image features. On the other hand, Wavelets have been used in detecting salient points representing local properties of images in content-based image retrieval, which is proposed by *Sebe et al.* [1]. *Schneiderman et al*. [52] describe a statistical method for 3D object detection by histogramming the subset of Wavelet coefficients and their position on the object. *Gonzalez-Audicana et al.* [53] use the multiresolution Wavelet decomposition to execute the spatial detail information extraction for image fusion. The primary advantage of these transformation methods is the removal of redundancy between neighboring pixels [19], and it leads to uncorrelated transform coefficients which can be encoded independently. Among all the transform techniques, Wavelet Transform has the best performance by localizing in frequency and also in space. Therefore Wavelet Transforms are desirable to deal with signals which are localized in time or space (speech or imagery) [8, 22].

V. CORRELOGRAM OF ORIENTED GRADIENT

Histogram is easy to compute and invariant to rotation and translation of the image content due to its capture of global features. Particularly, Histogramming the gradient information of images has been widely used in the past decades in image retrieval and indexing [42, 43, 44]. The principal idea behind these algorithms is that appearance and shape of local objects can be characterized well by the distribution of local intensity of gradients, even without the position knowledge of each gradient. Lowe's *Scale Invariant Feature Transformation (SIFT)* [42] is a milestone of making use of gradient information.

It is to extract and match scale invariant keypoints by building sub-region descriptors and adopting the local spatial gradient histogramming and normalization. Correlogram has been long used in computer vision and image retrieval areas. *Julesz* [46] uses gray level spatial dependence and *Haralick* [47] proposes to describe twodimensional spatial dependence of gray values by a multi-dimensional co-occurrence matrix. In [48], *Huang et al.* refine the co-occurrence matrix by maintaining only the correlation of pairs of color values as a function of distance. This statistic matrix captures the relationship between the spatial correlations of all possible pairs of features as a function of distance, while histogram only captures the global distribution of features from images. *Savarese et al.* [54] propose to use the correlogram for capturing the spatial arrangement of pixel labels (visual words) as a representation of an image. Again, *Savarese et al.* [49] extend correlogram to the spatial temporal domain.

TABLE 1 | Summary of previous surveys.



Authors and year	Area of interest
Aggarwal and Cai (1999)	Human motion analysis and tracking from single and multiview data
Gavrila (1999)	Shape model analysis from 2D and 3D data
Pantic and Rothkrantz (2003)	Multimodal human affective state recognition
Wang et al. (2003)	Human detection, tracking, and activity recognition
Moeslund et al. (2006)	Motion initialization, tracking, pose estimation, and recognition
Pantic et al. (2006)	Investigation of affective and social behaviors for human-computer interactions
Jaimes and Sebe (2007)	Multimodal affective interaction analysis for human-computer interactions
Turaga et al. (2008)	Categorization of actions and activities according to their complexity
Zeng et al. (2009)	Audio-visual affective recognition analysis
Poppe (2010)	Action classification according to global or local representation of data
Aggarwal and Ryoo (2011)	Gestures, human activities, actions, and interactions analysis
Bousmalis et al. (2013a)	Audio-visual behavior analysis of spontaneous agreements and disagreements
Chen et al. (2013b)	Human body part motion analysis from depth image data
Ye et al. (2013)	Human activity analysis from skeletal poses using depth data
Aggarwal and Xia (2014)	Human activity analysis from stereo, motion capture, and depth sensors 3D data
Guo and Lai (2014)	Understanding human activities from still images
Rodriguez et al. (2014)	Representation of human behavior ontologies from knowledge-based techniques

VI. CONCLUSION

Human Activity Recognition (HAR) is the understanding of human behaviour from data captured by pervasive sensors, such as cameras or wearable devices. It is a powerful tool in medical application areas, where consistent and continuous patient monitoring can be insightful. Wearable devices provide an unobtrusive platform for such monitoring, and due to their increasing market penetration, feel intrinsic to the user. This daily integration into a user's life is crucial for increasing the understanding of overall human health and wellbeing. This is referred to as the "quantified self" movement. Wearables, such as actigraph accelerometers, generate a continuous time series of a person's daily physical exertion and rest. This ubiquitous monitoring presents substantial amounts of data, which can (i) provide new insights by enriching the feature set in health studies, and (ii) enhance the personalization and effectiveness of health, wellness, and fitness applications. By decomposing an accelerometer's time series into distinctive activity modes or actions, a comprehensive understanding of an individual's daily physical activity can be inferred. The advantages of longitudinal data are however complemented by the potential of noise in data collection from an uncontrolled environment. Therefore, the data sensitivity calls for robust automated evaluation procedures.

In this paper, we present a robust automated human activity recognition (RAHAR) algorithm. We test our algorithm in the application area of sleep science by providing a novel framework for evaluating sleep quality and examining the correlation between the aforementioned and an individual's physical activity. Even though we evaluate the performance of the proposed HAR algorithm on sleep analysis, RAHAR can be employed in other research areas such as obesity, diabetes, and cardiac diseases.

VII. REFERENCES

[1] A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In Proceedings of IEEE International Conference on Computer Vision. pp. 726–733, 2013.

[2] A. Yilmaz and M. Shah. Recognizing human actions in videos acquired by uncalibrated moving cameras. In Proceedings of IEEE International Conference on Computer Vision. Vol. 1, pp. 150-157, 2005.

[3] Y. Song, L. Goncalves, and P. Perona. Unsupervised learning of human motion. IEEE Transactionson Pattern Analysis and Machine Intelligence. Vol. 25, pp. 814 - 827, 2003.

[4] M. Antonini, M. Barlaud, P. Mathieu, I. Daubechies. Image coding using wavelet transform. IEEE Transactions on Image Processing. Vol. 1, pp. 205-220.1992.

[5] N. Ahmed, T. Natarajan, K. R. Rao. Discrete Cosine Transfom. IEEE Transactions on Computers. Vol. C-23, pp. 90-93, 1974.

[6] E. O. Brigham, C. K. Yuen. The fast Fourier transform. IEEE Transactions on Systems, Man and Cybernetics. Vol. 8, pp. 146-146, 1978.

[7] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld. Learning realistic human actions from movies. IEEE Conference on Computer Vision and Pattern Recognition. pp.1-8, 2008. Website: http://www.irisa.fr/vista/Equipe/People/Laptev/download.html

[8] C.-C. Chang, C.-J.Lin. LIBSVM: a library for support vector machines, 2001. Software available at : http://www.csie.ntu.edu.tw/~cjlin

[9] N. Cristianini and J. Shawe-Taylor, An introduction to Support Vector Machines. Cambridge University Press, Cambridge. 2000.

[10] F. Suard, A. Rakotomamonjy, A. Bensrhair. Object categorization using kernels combining graphs and histograms of gradients. International Conference on Image Analysis and Recognition. Vol. 2, pp. 23–34, 2006.

[11] T. Kanungo, D. M. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A. Y. Wu. A local search approximation algorithm for k-means clustering. Computational Geometry: Theory and Applications. Vol. 28, pp. 89-112, 2004.

[12] R. C. Veltkamp, M. Hagedoorn. State of art in shape matching. In Principles of Visual Information Retrieval, M. S. Lew, Ed. Springer-Verlag, London. pp. 87-119, 2001.

[13] . Vetter, M. Jones and T. Poggio. A bootstrapping algorithm for learning linear models of object classes. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 40-46, 1997.

[14] D. Sharvit, J. Chan, H. Tek and B. B. Kimia. Symmetry-based indexing of image databases. Journal of Visual Communication and Image Representation. Vol. 9(4), pp. 366-380, 1998.

[15] Y. Gdalyahu, D. Weinshall. Flexible syntactic matching of curves and its application to automatic hierarchical classification of silhouettes. IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol 21(12), pp. 1312-1328, 2012.

CITE AN ARTICLE

Batra, V. K., & Gaur, P., Mrs. (2018). REVIEW PAPER ON REAL TIME HUMAN ACTION RECOGNITION USING SHAPE FEATURES HISTOGRAMS. *INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY, 7*(5), 418-422.