Pose-invariant kinematic features for action recognition

Ramanathan, Manoj; Yau, Wei-Yun; Teoh, Eam Khwang; Thalmann, Nadia Magnenat

2018

Ramanathan, M., Yau, W.-Y., Teoh, E. K., & Thalmann, N. M. (2017). Pose-invariant kinematic features for action recognition. Proceedings of the 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 292-299. doi:10.1109/APSIPA.2017.8282038

https://hdl.handle.net/10356/138068

https://doi.org/10.1109/APSIPA.2017.8282038

Pose-Invariant Kinematic Features for Action Recognition

Manoj Ramanathan*†, Wei-Yun, Yau†, Eam Khwang, Teoh* and Nadia Magnenat Thalmann‡
*School of Electrical and Electronic Engineering
Nanyang Technological University, Singapore
Email: manoj005@e.ntu.edu.sg, eekteoh@ntu.edu.sg
†Institute of Infocomm Research, A-STAR, Singapore
Email: wyyau@i2r.a-star.edu.sg
‡Institute of Media Innovation, Nanyang Technological University, Singapore
Email: nadiathalmann@ntu.edu.sg

***Abstract*—Recognition of actions from videos is a difficult task due to several factors like dynamic backgrounds, occlusion, pose-variations observed. To tackle the pose variation problem, we propose a simple method based on a novel set of pose-invariant kinematic features which are encoded in a human body centric space. The proposed framework begins with detection of neck point, which will serve as a origin of body centric space. We propose a deep learning based classifier to detect neck point based on the output of fully connected network layer. With the help of the detected neck, propagation mechanism is proposed to divide the foreground region into head, torso and leg grids. The motion observed in each of these body part grids are represented using a set of pose-invariant kinematic features. These features represent motion of foreground or body region with respect to the detected neck point's motion and encoded based on view in a human body centric space. Based on these features, pose-invariant action recognition can be achieved. Due to the body centric space is used, non-upright human posture actions can also be handled easily. To test its effectiveness in non-upright human postures in actions, a new dataset is introduced with 8 non-upright actions performed by 35 subjects in 3 different views. Experiments have been conducted on benchmark and newly proposed non-upright action dataset to identify limitations and get insights on the proposed framework.**

***Index Terms*—Action recognition, pose-invariance, kinematic features, human body centric space**

I. Introduction

With advent of new age cameras and a huge corpus of visual data has opened up research potential in visual processing and understanding. By recognizing actions of a person, critical and vital clues can be obtained for behavioural analysis. Also action recognition has several applications in surveillance, health care, physiotherapy etc making it a very hot topic of research in computer vision community. In recent years, researchers have proposed several methods to identify the action. We refer readers to look at action recognition survey for more information on these methods [1], [2]. As pointed out in [3], recognizing actions is not easy due to several bottlenecks such as view-invariance, occlusion, camera motion, anthropometric variations etc. One main bottleneck is the amount of pose-variations exhibited by the subject performing the action. Also actions can be performed in non-upright posture making it difficult to recognize them. In this paper, we propose a novel approach to represent the action using pose-invariant kinematic features in a human body centric space.

The framework aims to extract pose-invariant kinematic motion features by quantifying motion of body parts with respect to the body space itself. The framework consists of 4 main steps, namely, neck detection, body part grid division, extraction of kinematic motion features and encoding them in a newly proposed human-body centric space to compute pose-invariant features. Recently, [4] introduced a emotion recognition framework based on 4 kinematic features, namely, divergence, curl, projection and rotation, which measure motion of facial muscles with respect to fixed nose reference point. In this paper, we have chosen neck as a reference point to represent the body part's motion as it can be considered as stable and less susceptible to occlusions. For automatic neck detection, we introduce an algorithm that uses the output of the fully connected layer of a deep learning model as features.

Once the neck is detected, the above mentioned pose-invariant kinematic features can be measured. Projection and rotation features measure the motion of foreground region with respect to the identified neck point's motion. To quantify motion of each body part, we propose a propagation mechanism that divides the foreground region into 3 grids corresponding to head, torso and legs respectively. Using the detected neck and foreground, the mechanism estimates an approximate head size and uses it to propagate in the bodys orientation to identify the required grids.

The encoding and aligning of these features in the human-body centric space is based on the assumption that the person's viewing direction is known. The proposed framework is generalized to handle different body orientations. To test the ability to handle different body postures, we introduce a new non-upright action dataset (NUAD), containing 8 actions performed by 35 subjects. The dataset is captured in 3 different views.

The rest of the paper is organized as follows: Section II focuses on related work in the field of action recognition. In section III, we elaborate the proposed framework, comprising of deep learning based neck detection, computation of kinematic features, action representation in human-body centric space. Section IV provides details on NUAD dataset. Experimental results, limitations and discussion are discussed in section V. Conclusions and future directions are given in section VI.

II. Related work

Motion of various body parts characterize an action. Due to this several researchers use motion as the basis of their action representation. [5] proposed Motion History Images (MHIs) and Motion Energy Images (MEIs), which serve as a simple temporal template of motion occurring in the foreground. [6] extended MHI and MVI to volume based representation to allow for view point independent representation. [7], [8] proposed variants based on MHI and MVI. Motion trajectories [9], [10] aim to capture continuous movement of parts during an action. Optical flow is the simplest way of capturing motion and forms the basis of action representation in [11]–[13].

Kinematic features focus more on the dynamics of motion and can provide discriminative representation [14]. Due to this, [14], [15] used them as basis for recognizing actions. [14] proposed several kinematic features including divergence, vorticity, symmetric and anti-symmetric flow fields, second and third principal invariants of flow gradient and rate of strain tensor and third principal invariant of rate of rotation tensor. Each kinematic feature gives rise to a spatio-temporal pattern. Motion primitives were obtained in the form of kinematic modes. [15] introduced a new kinematic feature DCS descriptor (div, curl, shear). For the final representation, all possible pairings of kinematic features (div-curl, curl-shear and div-shear) were considered. [9] introduced MBH features which are computed based on horizontal and vertical derivatives of optical flow. Most methods represent motion in the spatial or image dimension. In this paper, we introduce a human-body centric space to encode motion of body parts with respect to the body itself.

In order to represent the motion of body parts, [16]–[18] extract STIP and represent the motion of the extracted points. [19] proposed motionlets, a mid-level spatio-temporal part, which are a tight cluster in motion and appearance space corresponding to each body part movements. As all these methods have tight integration with the pose, achieving pose-invariance is not possible unless multiple views of the same pose are captured as well. Deep learning methods [20]–[24] for action recognition try to learn both appearance and context. But two stream architectures fail to register appearance with optical flow and how these cues evolve over time [22]. In this paper, we introduce a new kinematic features (Proj and Rot) borrowed from emotion recognition [4] to represent motion of body parts with respect to neck point's motion. Also, the proposed method is generalized to handle non-upright postures as well. By representing motion of body parts with respect to the body itself allows for pose-invariance.

III. Proposed Framework

Actions can be considered as specific movements of body parts and arranged in a specific temporal order with the aim of performing some tasks. One main bottleneck in action recognition is the wide variety pose variations observed in a video. Research on non-upright human posture action recognition is very less and limited. To handle these challenges, we propose a pose-invariant action recognition framework that is generalized to all human postures. Motion based methods are usually restricted since they represent action in image dimension. In our method, we represent motion of body parts in a human body centric space, whose origin is the neck point of the person.

Figure 1 shows the proposed framework for action recognition. The input video is preprocessed to determine foreground region in the frames. The framework consists of 4 main stages, namely, neck detection, propagation mechanism to divide foreground into different body parts, kinematic features extraction and encoding the features in a human-body centric space for achieving pose-invariance. For neck detection, deep learning is employed on the foreground region. In the following subsections, above mentioned 4 stages are explained in detail.

A. Deep Learning based Automatic Neck Detection

In recent years, deep learning has had significant impact on machine learning and computer vision. For image related applications, deep learning has been widely used such as pose estimation [25], [26], image classification [27] and object recognition [28], [29]. For our framework, we intend to use deep learning for automatic detection of person's neck in the frame from the detected foreground. The neck will serve as the origin of human body centric space and to calculate pose-invariant kinematic features. The neck point is chosen because it is very stable and less susceptible to be occluded. We will explain the model training and neck detection algorithm next.

Detection of neck in images is not a simple task. Neck is relatively a small part compared to other body parts and is much difficult to detect in images due to change in view, occlusion etc. For training the classifier, we require huge number of positive and negative class images. In this paper, we have 30,000 images to train the classifier (5,000 positive and 25,000 negative images). Images for negative class comprise those of other body parts, different backgrounds and other objects. Figure 2 shows some sample set of positive and negative images collected by us for training. During the training stage, each image is passed through a pre-trained imagenet model [30] (imagenet-vgg-f model which is similar to [28]) and output of FC7 (fully connected network) layer is extracted as feature vector. The size of the feature vector is 4096×1. Using the extracted features, ELM [31] classifier is trained and used for neck detection in the video frames. We conducted a 3-fold cross validation on ELM to tune its parameters. A performance of 99.8% was observed during the cross-validation.

Once the foreground has been extracted, we apply a RMS error based segmentation [32] using x ,y components of optical flow and distance between pixels as features. Each of the identified segment is passed through the pre-trained imagenet

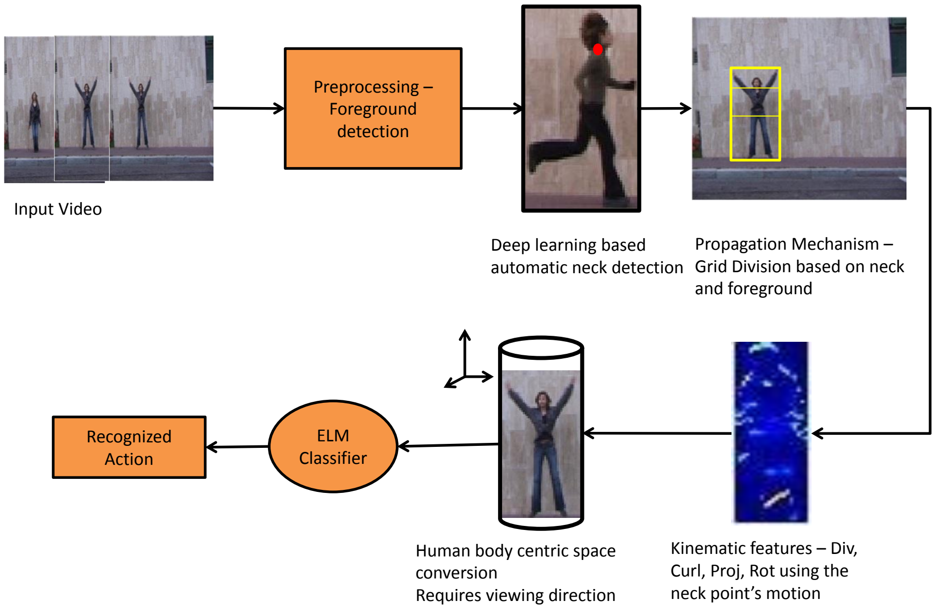
Input Video
Preprocessing – Foreground detection
Deep learning based automatic neck detection
Propagation Mechanism – Grid Division based on neck and foreground
Kinematic features – Div, Curl, Proj, Rot using the neck point's motion
Human body centric space conversion
Requires viewing direction
ELM Classifier
Recognized Action

Fig. 1. Proposed pose-invariant action recognition framework. Once foreground is detected, an automatic neck detection module based on deep learning is applied. Based on neck point and foreground, body is divided into head, torso, leg grids and kinematic features are computed and re-aligned in the human body centric space. These features represent motion of the body parts with respect to the body itself thus allowing for pose-invariance.

model [30] and output of FC7 layer is extracted as features. The ELM outputs positive/ negative class and certainty value for each segment. The positive segment with highest certainty value is chosen as the neck segment. The midpoint of the foreground pixels in this segment is detected as final neck point. Positional constraints can also be implied based on the neck detected in the previous frame. Figure 3 some of the neck points detected for sample images. In our experiments, we have observed that neck detection is good even in non-upright postures such as bend, pushup etc.

B. Propagation Mechanism

The aim of proposed propagation mechanism to divide the foreground region in to head, torso and head grids based on the detected neck and foreground. By dividing into various body part grids, we can obtain a discriminative representation for each body part's motion during the action. The mechanism proposed is based on the natural proportionality of human body. As shown in figure 4, if the person's head size can be estimated as $2x$, then the torso and legs size would be approximately $5x$ and $6x$ in size in the normal upright posture. Therefore, by estimating the head size, we can identify the other grids. The first step would be estimate the person's body orientation. The direction between detected neck point and center of mass (from foreground), can be used as the body orientation. Using the body orientation and neck point the head grid's location and size can be estimated. Our propagation mechanism uses this approximate head size and propagates in the direction of body orientation to obtain head, torso and leg grids. This mechanism is able to identify grids correctly even for non-upright postures as long as the segmentation is correct.

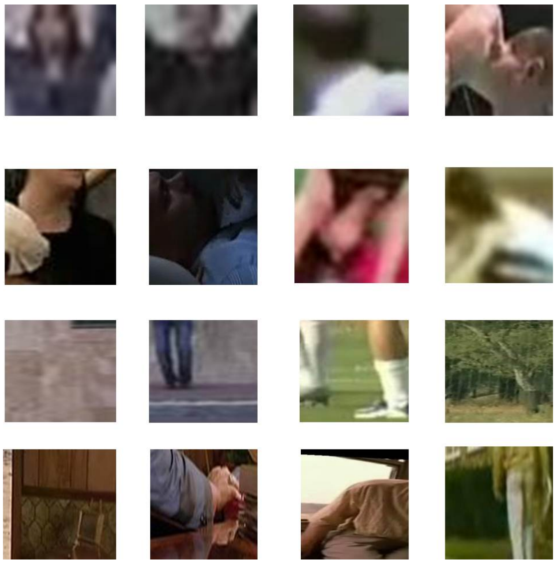

Fig. 2. Sample training images collected for deep learning neck detection. Top two rows show positive images and bottom two rows show negative images.

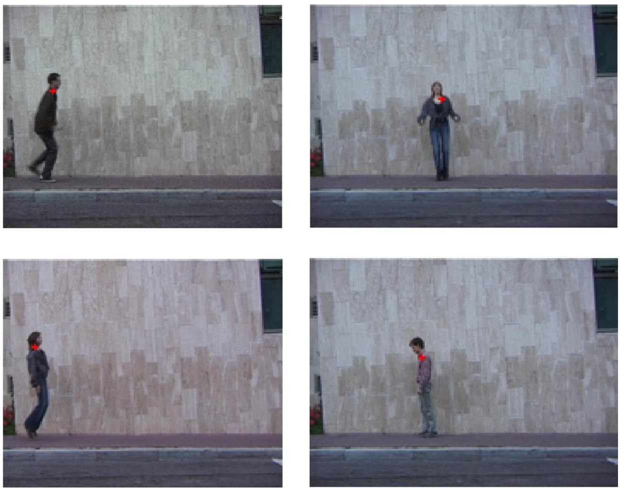

Fig. 3. Neck point detected using proposed algorithm. Neck is shown in red color in each frame.

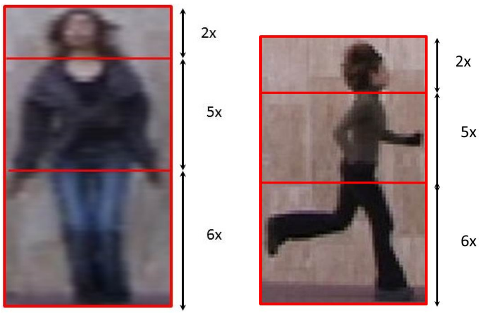

Fig. 4. (Top row) Proportionality of body parts used in Propagation Mechanism. Sample frames from Weizmann [33]

C. Kinematic Feature Extraction

The second stage involves computation of kinematic features, which contain more information on flow and dynamics of an action. The proposed features aim to compute and represent motion with respect to the body itself thereby resulting in a pose-invariant representation. A new set of kinematic features are proposed, namely, divergence, curl, Projection and Rotation respectively which is motivated from [4] used for emotion recognition. These features are described by the following equations

$$Div(p) = \frac{\partial OF_x(p)}{\partial x} + \frac{\partial OF_y(p)}{\partial y} \tag{1}$$

$$Curl(p) = \frac{\partial OF_y(p)}{\partial x} - \frac{\partial OF_x(p)}{\partial y} \tag{2}$$

$$Proj(p) = O\vec{F}_p \cdot \hat{P}_{neck} \tag{3}$$

$$Rot(p) = \hat{P}_{neck} \times O\vec{F}_p \tag{4}$$

where div measures the expansion and $curl$ measures the dynamics of circular motion at foreground pixel p in a local neighbourhood using OF_x and OF_y, the x and y components of optical flow at foreground pixel p respectively. $Proj(p)$ and $Rot(p)$ are two new kinematic features that measures the scalar product and vector product of OF_p (optical flow at p) with respect to a unit vector at a stable reference point. We used detected neck point ($\hat{P}_{neck}$) as the reference point in our current framework since it is stable and can be viewed easily from all directions.

D. Human Body Centric Space

The proposed human body centric space serves two purposes. Firstly, it adds to the pose-invariance nature of the proposed kinematic features. Secondly, [4] used un-weighted (only sign) and weighted (magnitude and sign) histograms of each kinematic feature separately accumulated over the entire video for emotion recognition. But it assumed that the face is frontal and so deal with only 2D motion of facial muscles. To account 3D motion of body parts by taking into account the different possible views, we introduce the human body centric space. We assume that the person's view is available to us. The body centric space can be thought of as cylinder with neck as the origin with three dimensions of the body namely, up-down, left-right and front-back as shown in figure 5(a). Given any frame, say figure 5(b), we can determine only two of these dimensions using the body orientation and the person's viewing direction. Therefore, divergence and projection will lie in one of these dimensions whereas curl and rotation will lie in the front-back dimension. Computed kinematic features will lie in either one of the dimensions of human-body centric space.

Based on the body orientation and viewing direction, the grids are divided into smaller cells as shown in figure 5(b). This allows us to capture more accurate representation within each body part grid. A $N \times 2$ cell configuration is employed for each grid. By dividing the grids into 2 halves, the whole body can be divided into either up-down, left-right or forward-backward halves based on observed view and orientation of the person. In short, it will correspond to the medial axis of the body (between neck and hip). We have conducted experiments for different settings of parameter N. While dividing into cells based on body orientation and viewing direction, we account for view angle and posture variation by labelling the cells as shown in figure 5(c). The kinematic features in each labelled cell are re-oriented based on alignment and accumulated over the entire video. Such labelling also allows us to compare in a more view-invariant manner.

The kinematic features are grouped into two classes, one comprising of Divergence and Curl that measure the local dynamics and the other comprises of Projection and Rotation, which are measured with respect to the detected neck point. Our action representation contains 2 components

a. Weighted and Unweighted histograms of divergence and curl for each human-body centric space dimension of each cell in each grid.
b. Weighted and Unweighted histograms of projection and rotation measured with respect to the neck reference point for each human-body centric space dimension of each cell in each grid.

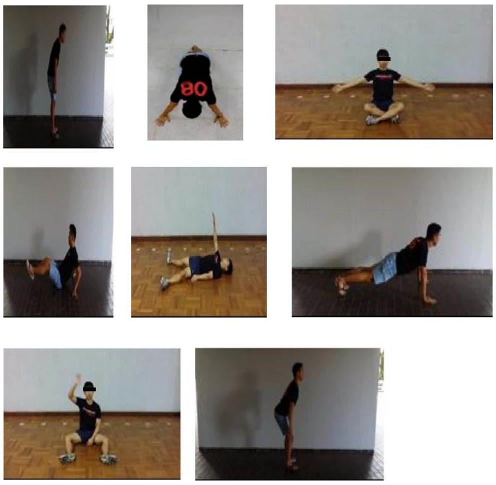

Fig. 6. Sample Images from NUAD dataset

Before histograms are computed, kinematic features for each labelled cell over the entire video are L2-normalized. This allow to account for inconsistencies that might arise due to speed variations. For action recognition, ELM [31] is trained with these histograms as features.

IV. Non-Upright Action Dataset (NUAD)

In this section, a new non-upright action recognition dataset is introduced. Currently traditional benchmark datasets available to the research community focus mainly on actions being performed in a upright standing posture. But the articulated nature of the body allows it to be in different postures. To develop a generalized pose-invariant action recognition dataset and explore the research study on non-upright postures, such a dataset is required. Therefore, a set of 8 actions have been chosen with a mix of standing, sitting and lying down poses included. The dataset contains 8 actions, namely, bending, climber, double-hand sitting wave, knee bender, pushup, single-hand sitting wave, lying down wave and squat. All these actions involve some kind of non-upright or sitting posture which is different from other datasets. Some of the example actions are shown in figure 6. The dataset consists 35 actors performing these 8 actions and captured from 3 views (front, left, right). The resolution of the videos is 192×108. Due to the non-upright posture, actions such as lying down wave, pushup, climber will be severely occluded in front view, therefore, these actions are captured from higher altitude (close to top view) such that most body parts are visible. To capture enough pose variations, action videos of 35 people are captured from 3 views. The dataset is captured in mostly uniform, simple backgrounds. The neck and foreground are available for the frames.

V. Experimental Results and Discussion

In this section, we explain in detail experiments conducted on benchmark datasets and identify limitations and insights on the proposed framework. The framework has been tested on 2 benchmark datasets Weizmann [33], KTH [34] and newly proposed NUAD dataset. The person's view is assumed to be available. For classification, we train the features with an ELM classifier with sigmoid activation function and number of hidden layer neurons are varied from $200 - 40000$ and the best performance is reported.

TABLE I
Performance Comparison for Weizmann dataset

Method	Performance (%)
Proposed Framework	100
[35]	100
[36]	100
[12]	99
[37]	98.9
[14]	95.75
[7]	93

TABLE II
Experiments for various cell configurations in Weizmann dataset

Cell Configuration	Proposed Framework Performance (%)
3 x 2	92.47
4 x 2	100
5 x 2	97.78

A. Weizmann Dataset

Weizmann dataset [33] contains 9 subjects performing 10 actions (bend, jack, jump, pjump, run, side, skip, walk, wave1, wave2) in a simple background and fixed cameras and evaluated in a leave one person out validation approach. Foreground masks are provided with the dataset and the proposed framework is applied on it. For dividing the grids into cells, we observed that 4×2 provides the best performance of 100%, which is comparable to state of the art methods. Our method is primarily based on kinematic features for action recognition that uses pose as a cue to quantify motion of each body part during an action. It can be noticed from table I that the proposed framework's performance is comparable to the state of the art methods that use both shape and motion features [35], [36].

We conduct an experiment on this dataset to evaluate for different $N \times 2$ cell configuration into which grids are sub divided into. The performance of the framework for 3 different cell configurations 3×2, 4×2 and 5×2 are tabulated in table II. We observed that 4×2 provides the best performance. For other cell configurations we observe the main confusion was between similar actions like *skip-jump*, *jump-walk* as body part motion cannot be captured accurately in these cell configurations. For other datasets we have used 3×2 cell configuration so that we do not over divide the body part grids. For other datasets, the variation in performance with different cell configurations is not very significant.

B. KTH Dataset

KTH dataset [34] comprises a total of 600 videos showing 6 actions (boxing, handclapping, handwaving, jogging, running,

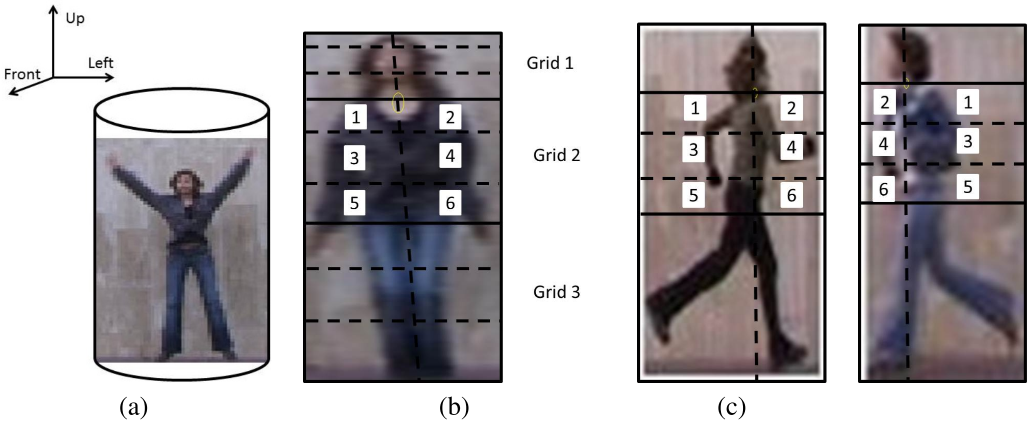

Fig. 5. (a) Human body centric space (b) Division of grids into smaller cells (c) Labelling of cells based on the change in view, body orientation, Only labelling of Grid 2 cells are shown

walking) performed by 25 actors in four different scenarios: outdoors (s1), outdoors with scale variation (s2), outdoors with different clothes (s3), and indoors (s4). KTH dataset is shot in relatively simple backgrounds with more or less stationary cameras except for zooming in and out. Using deformable parts model [38], bounding box is determined for foreground region. Average frame subtraction is applied within this bounding box to identify the foreground. The parts provided by [38] can also be used to identify the neck position in each frame. 3×2 cell configuration is used for feature computation.

KTH dataset is tested in two settings, 16 train + 9 test suggested split in [34] and leave one person out (LOV) setting. Results for both settings and comparison with other methods are tabulated in table III. It can be observed that under the first setting, our method provides a performance of only 87% which is less compared to methods using deep learning [39] and subspace analysis [40]. It might be noted that deep learning and subspace analysis would provide more discriminative power to their features. In the second setting, it can be noticed that several other methods have superior performance compared to ours. This can be attributed to subspace analysis [41], system based modelling approaches [41], [42], combining pose/shape with motion features [11], usage of contextual features [43]. Our method is primarily based on motion only and is better than methods that use kinematic features [14].

The six actions can be classified into 2 classes, hand actions (boxing, handclapping, handwaving) and leg actions (jogging, running, walking). The framework does not show any confusion between hand and leg actions. Considering the hand actions, boxing is performed in side view whereas handclapping and handwaving are done in a frontal view. From obtained results, boxing is identified 100% correctly and only confusion is between handclapping and handwaving. This indicates that by aligning the features on human-body centric space and encoding them based on the view is useful. It is noted that the waving action leads to incorrect head size estimate leading to wrong grids.

TABLE III
PERFORMANCE COMPARISON FOR KTH DATASET

Method	Performance (%)	
	16 + 9 [34]	LOV
Proposed Framework	87	90
[41]	-	99
[43]	-	96.16
[11]	-	95.5
[42]	-	95.17
[14]	-	87.7
[44]	89.34	93.18
[39]	94.39	-
[40]	93.6	-

Considering the leg actions, jogging, running and walking are very similar actions due to which erroneous detection are observed. There are two main reasons for such erroneous detection. No person is detected in several frames of video dataset by [38]. Due to this, motion of the person could not captured well to represent these action. Moreover to characterise and discriminate similar actions it would be necessary to have all the frames properly detected. The second drawback is the average background subtraction method. In many cases depending upon the background, the leg region of the foreground is not at all detected, which means leg motions cannot be learned properly. In addition, the presence of shadows also disrupt the neck detection and foreground detection. Due to the large number of actors, pose variations and execution rate variation exhibited is high. By identifying the pose, the recognition performance and pose-invariance of the framework can be improved.

C. Non-Upright Action Dataset (NUAD)

We introduce this new dataset to test the ability of proposed framework to handle non-upright postures and pose variations observed during an action. 3×2 cell configuration is used for feature computation. The foreground and neck point are provided with the dataset. The results of leave one person out setting performed on NUAD is tabulated in IV. The

TABLE IV
PERFORMANCE OF PMF PATH IN NUAD

Dataset	PMF path Performance (%)
NUAD	90.1

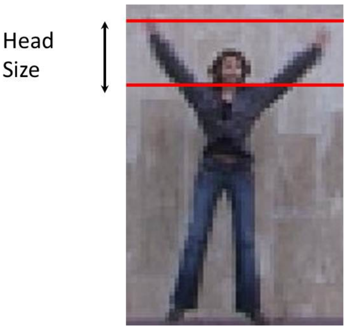

Fig. 7. Incorrect head size determined by propagation mechanism

proposed framework achieves a performance of 90.1% even in presence of non-upright postures indicating that it can handle such changes. As with previous datasets, confusion is observed between similar actions. For instance, there are 3 types of waving actions, *bend-squat*, *climber-pushup*. It is very difficult to differentiate these actions without knowledge of each body part. For instance, *bend-squat* appear quite similar and especially in front view.

D. Discussion

In the proposed framework, propagation mechanism divides the foreground into head, torso and leg grids. The head size, orientation and view determines the location of other grids. Consider the figure 7, the head size is wrong due to presence of hands. In these cases, the grid division will be wrong which in turn affects the final action representation. We observed this kind of error for *wave, climber, pushup etc* actions in the datasets. The head size in the action videos considered here are too small. A head or pose detection module can be included to obtain proper grid division. This is our future work

As observed in KTH dataset, foreground detection is another main reason for erroneous action recognition. This is one of the main reasons why it cannot be used for more difficult datasets such as HMDB51 [45]. For human body centric space, we assume that person's view point is available. View point can also be automatically estimated using cues such as body contour shape, head pose, major motion occurring direction etc. Also view point must be identified as one of four classes, namely, left, right, front and back. The proposed framework can only be applied for simple backgrounds. [46] can be used as a preprocessing stage to identify foreground regions in cluttered backgrounds. The proposed framework also shows confusion between similar actions. This is mainly because our framework is using only motion based features. For better performance, we need to include pose, appearance and contextual information. This will be a part of future work and extend it to more difficult benchmark datasets. Occlusion handling is also very essential for extension to difficult datasets. The performance in NUAD dataset shows our framework is generalized to handle non-upright postures as well as low resolution. The automated neck detection algorithm seems to work very well with all possible postures in the various datasets.

VI. CONCLUSION

Recognizing actions from videos is very difficult due to several bottlenecks such as view changes, occlusion, pose variations etc . Research in non-upright human posture action is also very limited. In this paper, we have proposed a simple framework for pose-invariant action recognition using a new set of kinematic features. These features represent motion of body parts with respect to the neck point's motion. For automatic neck detection, we introduce an algorithm based on deep learning features. The feature vector is obtained from fully connected layer of a pre-trained imagenet model. Also, we also introduce a new human body centric space to represent the motion with respect to the body itself, by assuming that the person's viewing direction is known. To test the effectiveness of the framework in non-upright human posture actions, a new dataset NUAD is introduced. This contains 8 actions performed by 35 actors in 3 views. The framework shows good performance in all human body postures and datasets. By including pose and appearance information, handling occlusions better performance can be achieved and be used for in the wild datasets.

ACKNOWLEDGEMENT

This research is supported by Institute of Infocomm Research, A-STAR, Singapore and School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. This research is partly supported by the Being Together Centre, a collaboration between Nanyang Technological University (NTU) Singapore and University of North Carolina (UNC) at Chapel Hill. The Being Together Centre is supported by the National Research Foundation, Prime Ministers Office, Singapore under its International Research Centres in Singapore Funding Initiative.

REFERENCES

[1] D. Weinland, R. Ronfard, and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," *Computer Vision and Image Understanding, Elsevier*, vol. 115, pp. 224 – 241, February 2011.

[2] M. Ryoo and J. Aggarwal, "Human activity analysis: A Review," *ACM Computing Surveys, Article 16*, vol. 43, pp. 16:1 – 16:43, April 2011.

[3] M. Ramanathan, W.-Y. Yau, and E. K. Teoh, "Human action recognition with video data: Research and evaluation challenges," *IEEE Trans. on Human Machine Systems*, vol. 44, pp. 650 – 663, October 2014.

[4] S. Shojaeilangari, W.-Y. Yau, K. Nandakumar, J. Li, and E. K. Teoh, "Robust representation and recognition of facial emotions using extreme sparse learning," *IEEE Trans. on Image Processing*, vol. 24, pp. 2140 – 2152, July 2015.

[5] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 257 – 267, March 2001.

[6] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Computer Vision Image Understanding*, vol. 104, pp. 249 – 257, October 2006.

[7] Y. Lu, Y. Li, Y. Chen, F. Ding, X. Wang, J. Hu, and S. Ding, "A Human action recognition method based on Tchebichef moment invariants and temporal templates," in *Intl. Conf. on Intelligent Human-Machine Systems and Cybernetics*, pp. 76–79, August 2012.

[8] M.-C. Roh, H.-K. Shin, and S.-W. Lee, "View-independent human action recognition with volume motion template on single stereo camera," *Pattern Recognition Letters, Elsevier*, vol. 31, pp. 639 – 647, May 2010.

[9] H. Wang, A. Klser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *Intl Journal on Computer Vision*, vol. 103, pp. 60 – 79, May 2013.

[10] Y. Chen, Z. Li, X. Guo, Y. Zhao, and A. Cai, "A spatio-temporal interest point detector based on vorticity for action recognition," in *IEEE Intl. Conf. on Multimedia and Expo Workshops*, pp. 1 – 6, July 2013.

[11] L. Liu, L. Shao, and P. Rockett, "Boosted key-frame selection and correlated pyramidal motion-feature representation for human action recognition," *Pattern Recognition 46, Elsevier*, pp. 1810 – 1818, July 2013.

[12] Y. Yang, I. Saleemi, and M. Shah, "Discovering motion primitives for unsupervised grouping and one-shot learning of human actions, gestures, and expressions," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 1635 – 1648, July 2013.

[13] K. Subramanian and S.Suresh, "Human action recognition using meta-cognitive neuro-fuzzy inference system," in *Intl. Joint Conf. on Neural Networks*, pp. 1 – 8, June 2012.

[14] S. Ali and M. Shah, "Human action recognition in videos using kinematic features and multiple instance learning," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 288 – 303, February 2010.

[15] M. Jain, H. Jegou, and P. Bouthemy, "Better exploiting motion for better action recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition*, June 2013.

[16] D. Liu, M.-L. Shyu, and G. Zhao, "Spatial-temporal motion information integration for action detection and recognition in non-static background," in *IEEE Intl. Conf. on Information Reuse and Integration*, August 2013.

[17] K. Hara, T. Hirayama, and K. Mase, "Simultaneous action recognition and localization based on multi-view hough voting," in *IAPR Asian Conf. on Pattern Recognition*, November 2013.

[18] A.-P. Ta, C. Wolf, G. Lavoue, A. Baskurt, and J.-M. Jolion, "Pairwise features for human action recognition," in *Intl. Conf. on Pattern Recognition*, pp. 3224 – 3227, August 2010.

[19] L. Wang, Y. Qiao, and X. Tang, "Motionlets mid-level 3d parts for human motion recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition*, June 2013.

[20] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *European Conf. on Computer Vision*, vol. 9912, pp. 20 – 36, October 2016.

[21] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems*, vol. 27, pp. 568 – 587, November 2014.

[22] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1933 – 1941, June 2016.

[23] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 4305 – 4314, June 2015.

[24] W. Zhu, J. Hu, G. Sun, X. Cao, and Y. Qiao, "A key volume mining deep framework for action recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1991 – 1999, June 2016.

[25] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1653 – 1660, June 2014.

[26] X. Liang, S. Liu, X. Shen, J. Yang, L. Liu, J. Dong, L. Lin, and S. Yan, "Deep human parsing with active template regression," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2015.

[27] S. Zhou, Q. Chen, and X. Wang, "Discriminative deep belief networks for image classification," in *Intl. Conf. on Image Processing*, pp. 1561 – 1564, September 2010.

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25* (F. Pereira, C. Burges, L. Bottou, and K. Weinberger, eds.), pp. 1097–1105, Curran Associates, Inc., 2012.

[29] D. Liang, K. Weng, C. Wang, G. Liang, H. Chen, and X. Wu, "A 3d object recognition and pose estimation system using deep learning method," in *Intl. Conf. on Information Science and Technology*, pp. 401 – 404, April 2014.

[30] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *British Machine Vision Conference*, September 2014.

[31] G.-B. Huang, L. Chen, and C.-K. Siew, "Universal approximation using incremental constructive feedforward networks with random hidden nodes," *IEEE Trans. on Neural Networks*, vol. 17, pp. 879 – 892, July 2006.

[32] M. Ramanathan, W.-Y. Yau, and E. K. Teoh, "Improving human body part detection using deep learning and motion consistency," in *IEEE Intl. Conf. on Control, Automation, Robotics and Vision*, pp. 1 – 5, November 2016.

[33] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Action as space-time shapes," in *IEEE Intl. Conf. on Computer Vision*, vol. 2, pp. 1395–1402, October 2005.

[34] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *Intl. Conf. on Pattern Recognition*, vol. 3, pp. 32–36, August 2004.

[35] Z. Jiang, Z. Lin, and L. S. Davis, "Recognizing human actions by learning and matching shape-motion prototype trees," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 34, pp. 533 – 547, March 2012.

[36] S. A. Rahman, S.Y.Cho, and M.K.H.Leung, "Recognising human actions by analysing negative spaces," *IET Computer Vision*, vol. 6, pp. 197 – 213, May 2012.

[37] T. Guha and R. K. Ward, "Learning sparse representations for human action recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 34, pp. 1576 – 1588, August 2012.

[38] I. Kokkinos, "Rapid deformable object detection using dual tree branch and bound," in *Advances in Neural Information Processing Systems 24*, pp. 2681 – 2689, December 2011.

[39] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Sequential deep learning for human action recognition," in *Human Behavior Understanding, Second International Workshop, Lecture Notes in Computer Science*, vol. 7065, Springer Berlin Heidelberg, November 2011.

[40] B. Li, M. Ayazoglu, T. Mao, O. I. Camps, and M. Sznaier, "Activity recognition using dynamic subspace angles," in *IEEE Conf. on Computer Vision and Pattern Recognition*, June 2013.

[41] M. T. Harandi, C. Sanderson, S. Shirazi, and B. C. Lovell, "Kernel analysis on grassmann manifolds for action recognition," *Pattern Recognition Letters*, vol. 34, pp. 1906 – 1913, November 2013.

[42] H. Wang, C. Yuan, G. Luo, W. Hu, and C. Sun, "Action recognition using linear dynamic systems," *Pattern Recognition 46, Elsevier*, pp. 1710 – 1718, June 2013.

[43] P. Bilinski and F. Bremond, "Contextual statistics of space-time ordered features for human action recognition," in *IEEE Intl. Conf. on Advanced Video and Signal-Based Surveillance*, pp. 228 – 233, September 2012.

[44] K. G. Derpanis, M. Sizintsev, K. J. Cannons, and R. P. Wildes, "Action spotting and recognition based on spatiotemporal orientation analysis," *IEEE Trans. On Pattern Analysis and Machine Intelligence*, vol. 35, pp. 527 – 540, March 2013.

[45] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: a large video database for human motion recognition," in *Proceedings of the International Conference on Computer Vision (ICCV)*, pp. 2556 – 2563, November 2011.

[46] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati, "Detecting moving objects, ghosts, and shadows in video streams," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 25, pp. 1337 – 1342, October 2003.