EasyChair Preprint

№ 4311

An Approach Towards Action Recognition using Part Based Hierarchical Fusion

Aditya Agarwal and Bipasha Sen

October 1, 2020

An Approach Towards Action Recognition using Part Based Hierarchical Fusion

Aditya Agarwal* and Bipasha Sen*

Microsoft India
{adiagar,bise}@microsoft.com

Abstract. The human body can be represented as an articulation of rigid and hinged joints which can be combined to form the parts of the body. Human actions can be thought of as a collective action of these parts. Hence, learning an effective spatio-temporal representation of the collective motion of these parts is key to action recognition. In this work, we propose an end-to-end pipeline for the task of human action recognition on video sequences using 2D joint trajectories estimated from a pose estimation framework. We use a Hierarchical Bidirectional Long Short Term Memory Network (HBLSTM) to model the spatio-temporal dependencies of the motion by fusing the pose based joint trajectories in a part based hierarchical fashion. To denote the effectiveness of our proposed approach, we compare its performance with six comparative architectures based on our model and also with several other methods on the widely used KTH and Weizmann action recognition datasets.

Keywords: Action Recognition · Hierarchical Bidirectional Long Short Term Memory Network · Part Based Fusion

1 Introduction

Human action recognition is a prominent field of research in computer vision with its wide applications in the areas of robotics, video search and retrieval, intelligent surveillance systems, automated driving, human computer interaction etc. Despite the extensive research on action recognition in the vision community, the problem still poses a significant challenge due to large variations and complexities, e.g., occlusion, appearance, low frame-rate, camera angle and motion, illumination, cluttered background, intra-class variations, and so on.

Traditionally, the spatio-temporal structure has been modeled using hand-crafted features and actions recognized using well defined discriminative networks. These methods usually start by detecting the spatio-temporal interest points [1] and then using local representations to describe these points. In [2], Histogram of Oriented Gradients (HOG)[3] and Histogram of Optical Flow (HOF)[4] were extracted at each spatio-temporal interest point and then features were encoded with Bag of Features (BoF). [5] proposed a method to extract dense

* these authors contributed equally to the work

trajectories by sampling and tracking dense points from each frame in multiple scales. They also extracted HOG, HOF and Motion Boundary Histogram (MBH) at each point whose combination further boosted the performance. In [6], dense trajectories were employed in a joint learning framework to simultaneously identify the spatial and temporal extents of the actions of interest in training videos. A combination of handcrafted and deeply learnt features was proposed in Trajectory-Pooled Deep-Convolutional Descriptors (TDD) [7] which proved to be successful. Deep Trajectory Descriptor (DTD) [8] for action recognition extracts dense trajectories from multiple consecutive frames and then projects them onto a two-dimensional plane to characterize the relative motion in frames. Despite encouraging results for action recognition on several datasets, these approaches suffer from variations of view point and scale, subject and appearance. Statistical and handcrafted features work well on recognizing simple actions but not complex actions involving multiple simultaneous sub-actions. Moreover, they are designed to be optimal for a specific task.

Recent advances in human pose estimation using deep learning [9–14] and the availability of depth sensors [15–18] have led to accurate representations of high level features. Studies show that high-level features extracted using current pose estimation algorithms already outperform state of the art low level representations based on hand crafted features implicating their potential in action recognition. In this work, we use a pose estimation framework based on Convolutional Neural Networks (CNN) in tandem with a robust object detection framework to deal with variations in scale and viewpoint to obtain a 2D representation of joint locations. The object detection algorithm filters frames that do not contain the object of interest and are therefore non-discriminative for the task of action recognition.

Human body can be articulated as a system of rigid and hinged joints [19]. These joints can be combined to form the limbs and the trunk. Human actions can be thought of as a collective action of these limbs and the trunk. Human action recognition is considered a time series problem where the characteristics of the body posture and its dynamics are extracted over time to represent the action [20–22]. [23] proposed a hierarchical approach on a trajectory of 2D skeleton joint coordinates. In this work, we propose a part based hierarchical action recognition pipeline on raw video sequences. We use a pose estimation framework to estimate a trajectory of 2D joint coordinates. These are combined in a part based hierarchical fashion using Hierarchical Bidirectional Long Short Term Memory (HBLSTM) networks to encode the spatio-temporal dependencies in the video sequence. The encoded representation is then fed to a discriminative network to classify the action.

The major contributions of this work are two-fold,

1. Designing a pipeline for pose based action recognition using a part-based hierarchical approach on raw video sequences. We handle the common issues of occlusion, camera zoom-motion-angle variations and eliminate non-discriminative frames while learning robust joint coordinates.

2. To validate the effectiveness of the proposed approach, we compare its performance with six other comparative architectures based on our model.

The rest of the paper is organized as follows, Section 2 defines our proposed pipeline and explains each component in detail. Section 3 talks about the experimental details and the evaluation results. Section 4 presents the conclusion and future work.

2 Proposed Approach

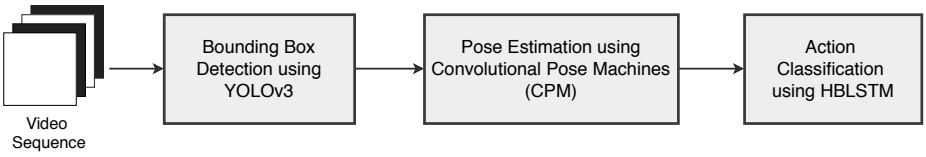

Fig. 1. Proposed pipeline

Fig. 1 depicts the proposed pipeline which consists of three modules that are connected sequentially:

1. Bounding Box Detection using YOLOv3 [24]
2. Pose Estimation using CPMs [12]
3. Action Classification using HBLSTM (Section 2.3)

A video sequence can be denoted as

$$F = [f_1, ..., f_T], f_t \in R^{160\times120\times1}, \tag{1}$$

where each frame f_t represents a single channel image. The proposed pipeline works in a supervised fashion. Frames of fixed dimension $\in R^{160\times120}$ (for Weizmann, dimension $\in R^{180\times144}$) are extracted from the input video sequence. An object detection algorithm is used to obtain an initial estimate of human's presence in the image. The purpose of the object detection module is two folds,

1. Generate accurate pose coordinates when dealing with changing camera zoom-motion-angles.
2. Select frames containing the human action and discard the remaining frames as they are non-discriminative for the task of action recognition.

We use a pose estimation framework based on CNNs to generate an estimate of human joints in the extracted bounding boxes. 2D coordinates of 14 joint locations (head, neck, wrist, elbow, shoulder, hip, knee, and ankle) are extracted for every frame representing the joint trajectories for the entire video sequence.

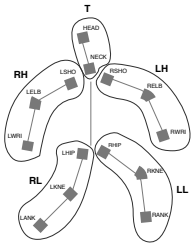

Fig. 2. Part representations of the joint coordinates.

The proposed model is based on HBLSTMs that takes 2D joint trajectories as input. HBLSTMs can learn across multiple levels of temporal hierarchy. As shown in Fig. 5., the first recurrent layer encodes the representation of 5 body parts, namely, Right Hand (RH), Left Hand (LH), Right Leg (RL), Left Leg (LL) and Trunk (T) (Fig. 2). The next set of layers encodes the part representations into Upper Right (RHT), Upper Left (LHT), Lower Right (RLT), and Lower Left (LLT) vectors by fusing the encoded representation of T with RH, LH, RL, and LL respectively. The subsequent layers generate the encoded representation of Upper (U) and Lower (L) bodies followed by the encoded representation of the entire body. Finally, a dense layer followed by a softmax layer is added to classify the action.

In the subsequent sections, we describe each of the modules in detail.

2.1 Bounding Box Detection

We use an object detection algorithm to obtain an initial estimate of human's presence in the image. YOLOv3 [24] is an efficient object detection algorithm pretrained on the ImageNet [25] and MSCOCO [26] datasets. It uses 53 successive 3×3 and 1×1 convolutional layers. The input to the bounding box algorithm is the grey scaled image frames of fixed dimension $\in R^{160 \times 120}$ (for Weizmann,

Fig. 3. The result of bounding box. (a) Shows the result on two frames with varying camera zoom. (b) Shows the result on the frame with missing person.

dimension $\in R^{180\times144}$) extracted from the video sequences in the KTH and Weizmann action recognition datasets. The network predicts 4 coordinates for each bounding box t_x, t_y, t_w and t_h. We have modified the code to produce bounding boxes only for those frames that have been labeled as a person and the remaining frames are discarded. Fig. 3 shows a typical output of the YOLOv3 algorithm. The reason behind using bounding box detection is two-fold,

1. It filters out all frames that do not contain human action and is non-discriminative for the task of action recognition.
2. It deals with camera zoom-motion-angle and generates accurate pose estimates.

2.2 Pose Estimation

Convolutional Pose Machines (CPMs) [12] were introduced for the task of articulated pose estimation. CPMs consist of a sequence of convolutional neural networks that repeatedly produce 2D belief maps for the location of each part. At each stage in a CPM, image features and belief maps produced in the previous stage are used as inputs producing increasingly refined locations of each part (Eq. 2). CPMs are based on pose inference machines [27] with the key difference being that prediction and image feature computation modules of a pose machine are replaced by deep convolutional architecture allowing for both image and contextual representations to be learned directly from the data.

$$g_t(x'_z, \psi_t(z, b_{t-1})) \rightarrow \{b_t^p(Y_p = z)\}_{p\in\{0\ldots P+1\}} \tag{2}$$

Fig. 4. Extracted Pose Coordinates using Convolutional Pose Machines

We operate CPMs directly on the bounding boxes generated from the previous stage to produce 2D joint coordinates. CPMs consist of five convolutional layers followed by two 1×1 convolutional layers and the input images being cropped to dimension 368×368. The bounding boxes extracted have a maximum dimension of 160×120 pixels for KTH and 180×144 pixels for Weizmann. Thus, the bounding boxes are first scaled while maintaining the aspect ratio and grey padded on either side to obtain the required dimension of 368×368. Fig. 4 shows a typical output generated by CPM on a single frame. A 14×2 representation is obtained for every frame which denotes the x, y coordinates of

the 14 joints in the frame. The coordinates of the frames are aggregated over the entire video sequence to obtain a joint trajectory of dimension $R^{T\times 28\times 1}$, T denoting the number of frames in the input video sequence.

2.3 Hierarchical Bidirectional Long Short Term Memory

We denote a video sequence as $X = [x_1, ..., x_T]$, with each frame x_t denoting the $2D$ coordinates of 14 joints. We work in a supervised classification setting with a training set,

$$\chi = \{(X_i, y_i)\}_{i=1}^{N} \in R^{T\times 28\times 1} \times \{1, ..., C\} \tag{3}$$

where X_i is a training video sequence and y_i is its class label (from one of the C possible classes).

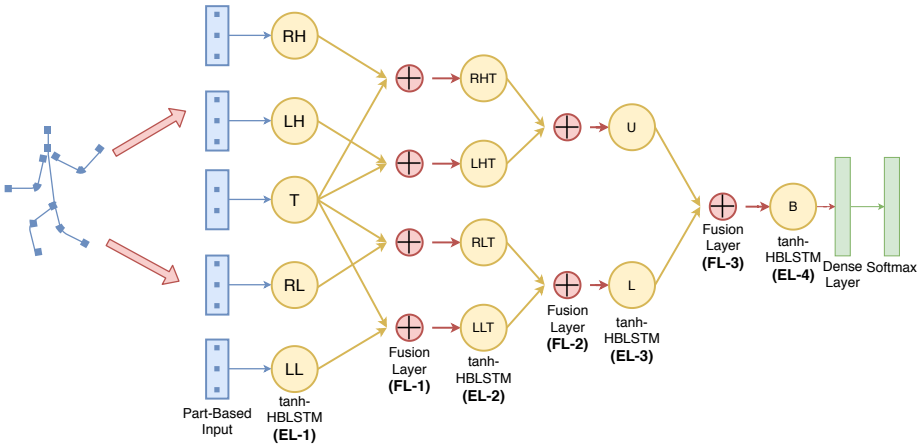

Fig. 5. Proposed Part Based Hierarchical Architecture. *EL* and *FL* denote *Encoding Layer* and *Fusion Layer* respectively.

Human body can be decomposed into five parts - two arms, two legs and a trunk (Fig. 2) and the global action can be modeled as the collective motion of these five parts. Benefitting from the LSTM's ability to model contextual dependencies from temporal sequences, we propose a Hierarchical Bidirectional LSTM (HBLSTM) for the task of pose based action recognition (Fig. 5). For the i^{th} encoding layer, given the inputs $I_{i,j}^t$ as trajectories of part j at i^{th} layer for time t, the corresponding encoded representation is expressed as

$$h_{i,j}^t = \overrightarrow{h_{i,j}^t} \oplus \overleftarrow{h_{i,j}^t} \tag{4}$$

where $\overrightarrow{h_{i,j}^t}$ and $\overleftarrow{h_{i,j}^t}$ are the forward and backward layers passes respectively with tanh activations [28].

For the fusion layer at time t, the newly fused p^{th} representation as the input for the $(i+1)^{th}$ encoding layer is:

$$I^t_{i+1,p} = h^t_{i,j} \oplus h^t_{i,k} \tag{5}$$

where $h^t_{i,j}$ is the concatenated hidden representation of the j^{th} part in i^{th} encoding layer and $h^t_{i,k}$ is for the k^{th} part in i^{th} layer.

The encoded representation of the entire body is given by the T^{th} unit of the last (4^{th}) encoding layer, $h^T_{4,body}$, which is given as input to the dense layer. The output of the dense layer is expressed as:

$$O = v_{h^t_{4,body}} \cdot h^T_{4,body} + b_{h^t_{4,body}} \tag{6}$$

O is given as input to the softmax layer. A softmax activation is applied to get the class probabilities as,

$$p(c_k) = \frac{e^{O_k}}{\sum_{j=1}^{C} e^{O_j}} \tag{7}$$

where C is the number of classes.

3 Experimental Details

3.1 Experimental dataset

As the proposed method aims to classify human actions in a video, we train the model on two of the most commonly used human actions dataset - KTH [29] and Weizmann [30] and evaluate its performance using the commonly used leave-one-out cross validation scheme based on the subjects. The data is prepared by sampling 25 consecutive frames of 2D joint coordinates estimated from Section 2.2 denoting a single human action $\in R^{25\times 28}$.

KTH dataset The KTH dataset contains six types of human actions: walking, jogging, running, boxing, hand-waving and hand-clapping performed by 25 subjects in four different scenarios: indoors, outdoors, outdoors with variations in scale, and outdoors with changes in clothing. The videos are on an average four seconds in duration with a static frame rate of $25fps$. Additionally, every sequence has an action that is performed 3 or 4 times, with a total of 2391 shorter subsequences. The sequences are taken over homogenous backgrounds and are sampled at a spatial resolution of 160×120 pixels.

Weizmann dataset The Weizmann dataset provided by [30] consists of 10 actions: bending, jumping, jumping jack, jumping in place, running, galloping sideways, skipping, walking, one-hand-waving and two-hands-waving. The dataset consists of 90 low resolution videos performed by 9 different subjects with a static frame rate of $25fps$ and are sampled at a spatial resolution of 180×144 pixels.

3.2 Class imbalance

We observed that for some of the video sequences belonging to the action recognition classes in the KTH dataset, the subject acting appears for a short duration of time in the video and thus the video is largely composed of frames that are non-discriminative for the given task. The number of frames in the KTH dataset per class for all video sequences is shown in Table 1. After preprocessing the frames using the object detection approach mentioned in Section 2.1, the number of bounding box frames for the classes jogging, running, and walking reduce drastically as the frames that don't contain the human as the object of interest are discarded. We notice that the number of frames for the remaining classes largely remains the same as can be seen from Table 1. We augment the dataset by adding a moving window of size 10 to handle the class imbalance problem. Table. 1 shows the number of sequences before and after data augmentation for each class.

Table 1. Class imbalance on the six action classes in the KTH dataset. *no-objdm* and *w-objdm* denote the number of frames before and after processing by object detection module. *no-daug* and *w-daug* denote the number of sequences before and after data augmentation.

	Frames		Sequences	
Classes	no-objdm	w-objdm	no-daug	w-daug
boxing	45187	45074	1848	**4401**
hand-clapping	42667	42448	1744	**4141**
hand-waving	53678	53291	2187	**5220**
jogging	43884	**18812**	800	**1779**
running	38505	**13001**	563	**1198**
walking	65795	32774	1343	**3122**

3.3 Origin shift of pose coordinates

The 2D pose coordinates estimated by the CPMs are anchored w.r.t. the top left position in the rectangular bounding box. Given that human actions are independent of their absolute spatial positions, we shift the pose coordinates with respect to the new origin at the center of the body. The new origin (O) is computed as the centroid of the pose coordinates belonging to head (P_{head}), left hip (P_{lhip}) and right hip (P_{rhip}) and is denoted as:

$$O = \frac{(P_{head} + P_{lhip} + P_{rhip})}{3} \tag{8}$$

The joint coordinates are shifted w.r.t the new origin as:

$$P'_{N,x}, P'_{N,y} = (P_{N,x}, P_{N,y}) - (O_x, O_y) \tag{9}$$

where P' and P denote the new and old pose coordinates respectively. Shifting the origin improved our action recognition rates by an average of $\sim 5\%$.

3.4 Experimental Results

Comparative Architectures To denote the effectiveness of our proposed architecture, we compare its performance with six other architectures based on deep RNN. Our first architecture is a Deep Bidirectional RNN (**DBRNN**) which is one of the most commonly used models in sequence classification problems. To denote the importance of LSTM units in modeling long term contextual dependencies and the role of backward connections in modeling the future context, we compare a Deep Unidirectional LSTM (**DULSTM**) and a Deep Bidirectional LSTM (**DBLSTM**). These architectures operate directly on the trajectories of 2D pose coordinates.

To denote the importance of hierarchical connections, we compare a Point based Hierarchical BLSTM model (**PointHBLSTM**) that first obtains an encoded representation of all the frames in a sequence and then encodes them temporally to obtain a representation of the entire sequence.

Table 2. Average recognition rates with different experiments on the augmented dataset from Section 3.2 using the leave-one-out cross validation scheme.

Methods	KTH	Weizman
DBRNN	82.4%	81.2%
DULSTM	89.8%	91.7%
DBLSTM	92.7%	94.8%
PointHBLSTM	94.1%	96.6%
PartHBLSTM$_1$	98.9%	99.9%
PartHBLSTM$_2$	98.4%	99.7%
Proposed Approach	**99.3%**	**100%**

To denote the effectiveness of part based hierarchical fusion, we build three similar part based models with different part based fusion. In the first part based model (**PartHBLSTM**$_1$), the encoded representation of upper body is obtained from the encoded representation of right hand, trunk and left hand and the lower body from encoded representation of left and right leg. The encoded representation of upper and lower bodies are then combined to get an encoded representation of the whole body. In the second part based model (**PartHBLSTM**$_2$), the encoded representation of left and right bodies are combined to obtain an encoded representation of the whole. Finally, we achieve the best evaluation result on the proposed approach shown in Fig. 5.

The number of learnable layers in all the experiments has been kept the same with small modifications to parameters to ensure best average performance.

Table 3. Recognition rates against different approaches on the KTH dataset.

Existing Methods	KTH	Weizman
Schuldt *et al.* [29]	71.72%	-
Fathi *et al.* [32]	90.5%	99.9%
Baccouche *et al.* [33]	94.39%	94.58%
Gorelick *et al.* [34]	-	97.83%
Gilbert *et al.* [35]	96.7%	-
Mona *et al.* [36]	97.89%	-
Proposed Approach	**99.3%**	**100%**

Evaluation on KTH and Wiezman Similar to many evaluation approaches on the KTH [29] and Weizmann [30] dataset, we carried out our experiments using the leave-one-out cross validation strategy [31] (i.e. all subjects except one were used for training and the learned model was evaluated on the remaining one) on the datasets described in Section 3.1. Average accuracy on the comparative architectures are reported in Table 2 and the average accuracy on the proposed model along with the existing approaches are reported in Table 3.

We achieve full separation between different actions and that a slight misclassification occurs between similar actions (i.e. between "hand-clapping" and "hand-waving" and between "running" and "jogging") in the KTH dataset. The classification accuracy is averaged over all selections of test data to achieve a recognition rate of **99.3%** and **100%** on the KTH and Weizmann dataset respectively using the proposed approach. We show that our proposed approach outperforms the state of the art on both KTH and Weizmann dataset for the task of action recognition.

4 Conclusion and Future Work

In this work, we present a pipeline for the task of human action recognition in videos. Using an object detection approach, we first estimate the presence of human and discard remaining frames as they are non-discriminative for the task of action recognition. We use a pose estimation framework to generate the trajectory of joint coordinates and combine them in a part-based hierarchical fashion to obtain a global representation of the entire video sequence. We showed that adding part based hierarchical fusion helps us achieve better results over other comparative architectures. Experimental evaluation on the KTH and Weizmann action recognition dataset show that our proposed approach outperforms the existing state of the art approaches on these datasets.

Future work includes applying the proposed system on more complex and larger datasets, such as UCF101 [37] and HMDB51 [38]. We are also exploring the use of appearance features in distinguishing between actions with similar motion but distinguishable appearance which is a limitation of the current approach.

References

1. Laptev, Lindeberg: Space-time interest points. In: Proceedings Ninth IEEE International Conference on Computer Vision. (2003) 432–439 vol.1
2. Poppe, R.: A survey on vision-based human action recognition. Image and Vision Computing **28** (2010) 976 – 990
3. Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In Leonardis, A., Bischof, H., Pinz, A., eds.: Computer Vision – ECCV 2006, Berlin, Heidelberg, Springer Berlin Heidelberg (2006) 428–441
4. Perš, J., Sulić, V., Kristan, M., Perše, M., Polanec, K., Kovačič, S.: Histograms of optical flow for efficient representation of body motion. Pattern Recognition Letters **31** (2010) 1369 – 1376
5. Wang, H., Kläser, A., Schmid, C., Liu, C.: Action recognition by dense trajectories. In: CVPR 2011. (2011) 3169–3176
6. Zhou, Z., shi, F., Wu, W.: Learning spatial and temporal extents of human actions for action detection. IEEE Transactions on Multimedia **17** (2015) 1–1
7. Wang, L., Qiao, Y., Tang, X.: Action recognition with trajectory-pooled deep-convolutional descriptors. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
8. Yemin Shi, Wei Zeng, Tiejun Huang, Yaowei Wang: Learning deep trajectory descriptor for action recognition in videos using deep neural networks. In: 2015 IEEE International Conference on Multimedia and Expo (ICME). (2015) 1–6
9. Toshev, A., Szegedy, C.: Deeppose: Human pose estimation via deep neural networks. 2014 IEEE Conference on Computer Vision and Pattern Recognition (2014)
10. Brau, E., Jiang, H.: 3d human pose estimation via deep learning from 2d annotations. In: 2016 Fourth International Conference on 3D Vision (3DV). (2016) 582–591
11. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation (2016)
12. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines (2016)
13. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields (2018)
14. Güler, R.A., Neverova, N., Kokkinos, I.: Densepose: Dense human pose estimation in the wild (2018)
15. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. In: CVPR 2011. (2011) 1297–1304
16. Ye, M., Zhang, Q., Wang, L., Zhu, J., Yang, R., Gall, J. In: A Survey on Human Motion Analysis from Depth Data. (2013) 149–187
17. Siddiqui, M., Medioni, G.: Human pose estimation from a single view point, real-time range sensor. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops. (2010) 1–8
18. Plagemann, C., Ganapathi, V., Koller, D., Thrun, S.: Real-time identification and localization of body parts from depth images. In: 2010 IEEE International Conference on Robotics and Automation. (2010) 3108–3113
19. Vemulapalli, R., Arrate, F., Chellappa, R.: Human action recognition by representing 3d skeletons as points in a lie group. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. (2014) 588–595

20. Gong, D., Medioni, G., Zhao, X.: Structured time series analysis for human action segmentation and recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **36** (2014) 1414–1427
21. Zhang, Y., Zhang, Y., Zhang, Z., Bao, J., Song, Y.: Human activity recognition based on time series analysis using u-net (2018)
22. Kim, H., Kim, I.: Human activity recognition as time-series analysis. Mathematical Problems in Engineering **2015** (2015) 1–9
23. Yong Du, Wang, W., Wang, L.: Hierarchical recurrent neural network for skeleton based action recognition. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2015) 1110–1118
24. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement (2018)
25. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR09. (2009)
26. Lin, T.Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollár, P.: Microsoft coco: Common objects in context (2014)
27. Ramakrishna, V., Munoz, D., Hebert, M., Bagnell, J., Sheikh, Y.: Pose machines: Articulated pose estimation via inference machines. (2014) 33–47
28. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing **45** (1997) 2673–2681
29. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local svm approach. In: Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004. Volume 3. (2004) 32–36 Vol.3
30. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: The Tenth IEEE International Conference on Computer Vision (ICCV'05). (2005) 1395–1402
31. Gao, Z., Chen, M.y., Hauptmann, A.G., Cai, A.: Comparing evaluation protocols on the kth dataset. In Salah, A.A., Gevers, T., Sebe, N., Vinciarelli, A., eds.: Human Behavior Understanding, Berlin, Heidelberg, Springer Berlin Heidelberg (2010) 88–100
32. Fathi, A., Mori, G.: Action recognition by learning mid-level motion features. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. (2008) 1–8
33. Baccouche, M., Mamalet, F., Wolf, C., Garcia, C., Baskurt, A.: Sequential deep learning for human action recognition. In Salah, A.A., Lepri, B., eds.: Human Behavior Understanding, Berlin, Heidelberg, Springer Berlin Heidelberg (2011) 29–39
34. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence **29** (2007) 2247–2253
35. Gilbert, A., Illingworth, J., Bowden, R.: Fast realistic multi-action recognition using mined dense spatio-temporal features. (2009) 925 – 931
36. Moussa, M.M., Hamayed, E., Fayek, M.B., El Nemr, H.A.: An enhanced method for human action recognition. Journal of Advanced Research **6** (2015) 163 – 169
37. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild (2012)
38. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: Hmdb: A large video database for human motion recognition. In: 2011 International Conference on Computer Vision. (2011) 2556–2563