Curvature: A signature for Action Recognition in Video Sequences

He Chen
Johns Hopkins University
hchen136@jhu.edu

Gregory S. Chirikjian
National University of Singapore, Johns Hopkins University
gchirik1@jhu.edu

Abstract

In this paper, a novel signature of human action recognition, namely the curvature of a video sequence, is introduced. In this way, the distribution of sequential data is modeled, which enables few-shot learning. Instead of depending on recognizing features within images, our algorithm views actions as sequences on the universal time scale across a whole sequence of images. The video sequence, viewed as a curve in pixel space, is aligned by reparameterization using the arclength of the curve in pixel space. Once such curvatures are obtained, statistical indexes are extracted and fed into a learning-based classifier. Overall, our method is simple but powerful. Preliminary experimental results show that our method is effective and achieves state-of-the-art performance in video-based human action recognition.

1. Introduction

Action Recognition based on AI-reasoning is one of the most important research topics in computer vision[6, 26, 37, 9]. In recent years, this field has witnessed big breakthroughs, and the research interest is evolving from learning the joints of human and recognizing human pose to the understanding of actions and scenes.

With the booming of learning-based AI technology, precision of action recognition has been raised to a new level[27]. However, challenges still exist in this field. Most of these learning-based algorithms are thirsty for large datasets with tens of thousands of videos and corresponding labels. Although thousands of videos are uploaded to Youtube every second, they are raw data which can't be fed directly into most of existing algorithms. The most time-consuming part is getting detailed, fine-tuned annotations, which usually involves some manual efforts[2, 41].

Encoding action into a model is important because this might lead to effective extraction of motion features, thus effectively decrease the amount of training data required. Modeling of human action could be categorized into two perspectives, namely spatially and temporally. Spatially,

the shape of human could be modeled into several forms such as skeleton[23], silhouettes[33], virtual skin[39], etc. Temporally, action could be modeled into video sequence, audio sequence, dynamic image[3], etc. In this paper, we focus on the second perspective.

In many application scenarios, actions or scenes need to be understood on universal time scale. For example, consider the scenario of service robot in a house as shown in Fig.1(a). By capturing a single image, the robot could be confused whether the lady wants to put down or pick up the cup. What if we capture six consecutive frames and see that the hand is moving towards table? Does she want to put the cup onto the table? Still hard to tell. It could be an old lady with trembling hand trying to pick up the cup. Locally, the cup could be approaching the table, while globally, the lady might want to pick up the cup from the table.

Figure 1. (a) Instance of action recognition in the application of service robots (b) Multimodal distribution

From the aspect of modeling, the problem of action recognition boils down to decreasing the distance between the distributions actions from the same class, while at the same time increasing the distance between that of different classes on universal time scale. In order to get a better understanding of sequences on a global timescale, [7] proposed the theory of globally optimal reparameterization algorithm (GORA) using variational calculus.

Inspired by [7], in this paper, we propose a novel few-shot learning algorithm using curvature as a signature for action recognition. In this algorithm, we use the curvature

of an optimally reparameterized video sequence as a signature of action. Such a model is based on viewing a video sequence as a curve in the pixel space, the curve parameters defining the speed of traversal through the sequence. The proposed algorithm considers and analyzes action from the global time scale, which is the arclength of the curvature, and expresses actions in a very compact way. After curvatures are obtained, we gather important features from these curvatures and input them into the classifier. Random forest is selected as the classifier. And it should be noted that our algorithm is robust against flipping, which is to say walking from left to right has the same signature as walking from right to left. This algorithm is also invariant to mirror reflections and rigid-body displacements of image plane. Moreover, because we modeled the distribution of data with an effective and compact model, our algorithm is not thirsty for large-scale data. In order to eliminate the noise of background, we used state-of-the-art segmentation algorithm Mask R-CNN as a pre-processing. Experimental results show that the proposed algorithm is effective and has good performance.

2. Related Work

In this section, a brief review is given for existing state-of-the-art methods in the field of action classification and representation. Due to limitation of space, we could not list all of them, but please refer to the reviews[24, 34] if you have further interest.

2.1. Action Classification Based on RGB Videos

In the early stage of the development of action recognition, histogram of gradient (HOG) and histogram of flow (HOG) algorithms are widely accepted structures. The main idea of HOG[8] is dividing images into small cells, and draw the histogram of the gradients of edges for these cells. While HOF[17] is based on measuring the angle between optical flow vector and x-axis. In[31, 32], H.Wang et al. proposed dense trajectory as a description of videos to accomplish action classification task, and improved it to be more robust against camera motions with optical flow. Thanks to the development of computer structure and learning theory, recent years have witnessed significant improvements of learning-based solutions for action recognition on RGB videos. S.Ji et al.[18] extended conventional CNN from image classification to videos by taking the time scale into consideration. And D.Tran et al.[10] made further improvement on CNN based method by modeling appearance and motion simultaneously and found the best kernel for the network. In [19], A.Karpathy et al. proposed Deep Video algorithm which enables CNN with the ability to classify large-scale dataset of 1 million videos with the multi-resolution architecture. Admittedly, these algorithms show excellent results on most benchmark datsets, the prerequisit of them usually includes a large training dataset.

2.2. Representation of Actions

In recent years, more and more attention is paid to alleviate the training burden by making improvements to modeling[35]. Girdhar at al. [13] learned attention maps instead of whole videos to make classification more efficient. In [15], Guo et al. proposed a graphical representation for spatial information in 3D data using neural graph matching networks, which enables few-shot learning. Bilen et al.[3] proposed a novel compact representation of actions using dynamic image. In their algorithm, dynamic images are obtained by directly applying rank pooling on the raw image pixels of a video, producing a single RGB image representation for each video. In [40], Zhou et al. used temporal relational reasoning to analyze the current situation relative to the past and formulate what may happen next. Long short term memory (LSTM) algorithm was introduced to cope with video sequences in the field of action recognition [11, 1]. LSTM, the well known language processing algorithm is breaking its way into the vision field, showing promising results and huge potential. Experimental results of multiple papers show that LSTM outperforms CNN-based methods when it comes to understanding action or scene. The underlying structure is natural because LSTM architecture models the input as a sequence, which is closer to the physics meaning of action.

3. Proposed Signature

For human, the task of recognition is one of the first skills we learn from the moment we are born and is one that comes naturally and effortlessly as adults. It is commonly believed that edge-detection of objects plays important role in the eyesight of human. However, the optical system of the human eye is vastly more complicated than edge-detector. Medical research carried out by L.Riggs in literature [22] shows that curvature is a specific feature of human vision perception, and detectors of curvature exist in human eyes that generate color-contingent aftereffects. In [20], M.Kass et.all proposed an active contour model called snake, which is also a very interesting way to describe curvature in the image plane. L.Gorman proposed in [14] that conversion of a image into a representation of curves and line features enable economical storage of information.

Unlike many works, we are not focusing on curvature of contours in the image plane, but rather curvature of the video sequence, which is viewed as a curve in d^2 dimensional Euclidean space. In this section, we firstly give a brief review of globally optimal reparameterization algorithm (GORA), then present the generation of curvature based on GORA.

3.1. Review of GORA

With the goal of comparing the similarity among video sequences, one important premise is to align the signals on temporal scale. [7] defined the arc length as a metric for signal alignment. In this paper, we depart from the arc length for reparameterization on temporal scale and push it further to use curvatures in high dimensional Euclidean space to classify actions.

Variational Calculus addresses the problem of seeking vector-valued functinos $\mathbf{x}(t)$ that minimizes the functional of the form

$$J = \int_0^1 f(x, \dot{x}, t)dt. \tag{1}$$

As proved in [7], when the integrand in Eqn. (1) is of the form

$$f(x, \dot{x}) = \dot{x}^2 g(x), \tag{2}$$

we can find the global minimum solution of Eqn. (1) by solving the Euler-Lagrange equations.

The construction of Eqn. (2) was arisen from the assumption that the metric space is comprised of matrix norm of the difference of two elements

$$d(X(t+dt), X(t)) = \left\lVert dX/dt \right\rVert dt, \tag{3}$$

where $X(t)$ is a matrix describing image pixel intensities w.r.t time in a video sequence. If we represent the reparameterized version of $X(t)$ as $Y(t) = X(\tau(t))$, then minimizing the integral over $t \in [0, 1]$ of

$$\left\lVert \frac{dY}{dt} \right\rVert^2 = \left\lVert \frac{dX(\tau(t))}{dt} \right\rVert^2 = \left\lVert X'(\tau) \right\rVert^2 \dot{\tau}^2 \tag{4}$$

gives the form of Eqn. (2).

To this light, assume that $X_1(t)$ and $X_2(t)$ are two signals on the time scale $t \in [0, 1]$, obtained by respectively vectorizing matrices $X_1(t)$ and $X_2(t)$, which might have different temporal frequencies over this range. Reparameterization is carried out by deriving a monotonically increasing function $\tau(t)$ that normalize signals onto universal standard timescale. The expression for $\tau(t)$ is derived using variational calculus, whose result is

$$\tau(t) = F^{-1}(t), \tag{5}$$

where F has the form of

$$F(\tau) = \frac{1}{c} \int_0^\tau \left\lVert \frac{dX}{dt} \right\rVert dt, \tag{6}$$

$$c = \int_0^1 \left\lVert \frac{dX}{dt} \right\rVert dt.$$

Then, if $X_i(t) = X(s_i(t))$, for $i = 1, 2$, where $s_i(t)$ are arbitrary monotonically increasing functions of time, reparameterization will recognize them as the same signal, by quotienting out the effects of $s_i(t)$.

3.2. Curvature Generation

Intuitively, curvature is the amount by which a geometric object, such as a surface, deviates from being a flat plane or a curve from being straight as in the case of a line[30].

That is the narrower definition of curvature. In this paper, we mean the broader definition of curvature, which is calculated by nth-order derivatives of a curve in n-dimensional Euclidean space. By curvature, we are referring to the curvature of the curve using the Frenet frame [30]. Assume that $\boldsymbol{X} = \boldsymbol{X}(s) \,\epsilon\, E^n$ is a parametric representation of a generally curved curve with arclength s. Then the following derivation equations are valid

$$\frac{d\boldsymbol{X}}{ds} = \boldsymbol{e}_1, \tag{7}$$

$$\frac{d\boldsymbol{e}_1}{ds} = \boldsymbol{e}_2 k_1, \tag{8}$$

$$\frac{d\boldsymbol{e}_2}{ds} = -\boldsymbol{e}_1 k_1 + \boldsymbol{e}_3 k_2, \tag{9}$$

$$......$$

$$\frac{d\boldsymbol{e}_{n-1}}{ds} = -\boldsymbol{e}_{n-2} k_{n-2} + \boldsymbol{e}_n k_{n-1}, \tag{10}$$

$$\frac{d\boldsymbol{e}_{n-1}}{ds} = -\boldsymbol{e}_n k_{n-1}, \tag{11}$$

where $s = F \cdot c$ represents the arc length, $\boldsymbol{e}_1, ..., \boldsymbol{e}_n$ represents the orthogonal basis, and $\boldsymbol{k}_1, ..., \boldsymbol{k}_n$ represents the curvatures. $\boldsymbol{k}_1$ is the curvature that is most widely used, while $\boldsymbol{k}_2, ..., \boldsymbol{k}_n$ are higher curvatures. Since curvature is an intrinsic quality, if a reflection, rotation, or translation is applied to the video sequence, the curvature will be invariant.

4. Action Recognition Based on Curvature Calculation

Now that we already compressed the patterns of actions into curvatures on the global timescale, we could use these curvatures to accomplish the action recognition task. In order to achieve that, first of all, consider the curvatures as one-dimensional distributions with respect to global time scale. Then, we input these features into a classifier to distinguish between different actions. Here, we choose to use Random Forest, which is one of the most powerful machine-learning based classifiers has a simple structure, very easy to implement, and usually wouldn't overfit.
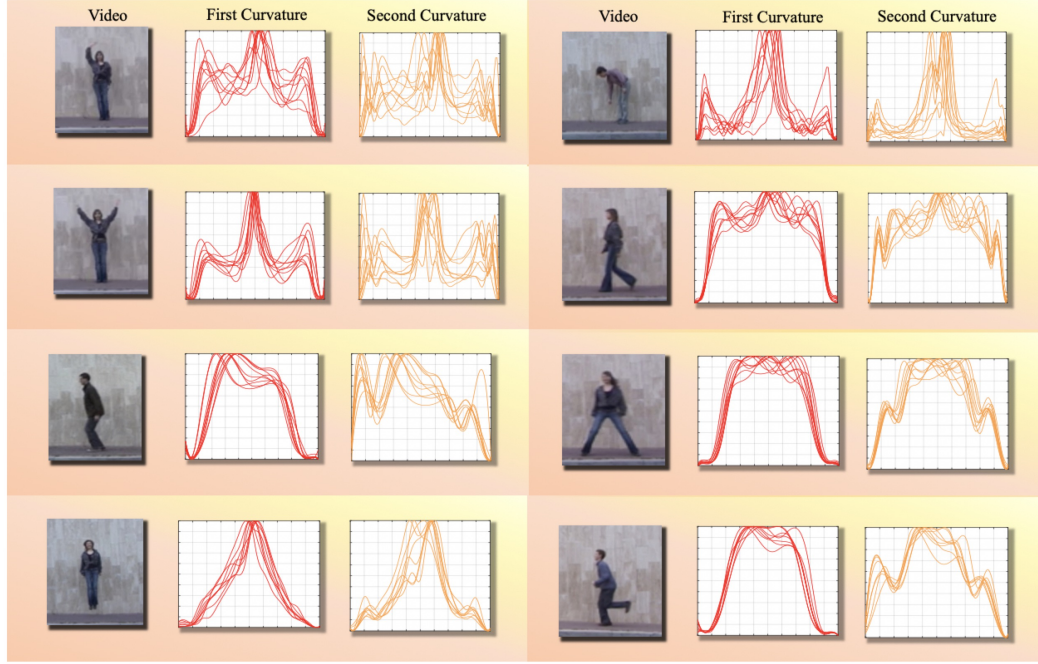
Figure 2. Curvature as a signature obtained on videos provided by Weizzman Dataset [4]

4.1. Feature Selection

When it comes to one-dimensional distributions, most commonly applied features including

- feature of center position: mean μ, median m

- feature of divergence: range r, standard deviation σ

And fancier statistical features including

- wave rate: ninty-percent-quantile minus ten-percent-quantile t

- skewness: a measure of asymmetry about a distribution about its mean

$$Skew(X) = E\left[(\frac{X-\mu}{\sigma})^3\right] \qquad (12)$$

- kurtosis: a measure of sharp or flat about a distribution

$$Kurt(X) = E\left[(\frac{X-\mu}{\sigma})^4\right] \qquad (13)$$

However, all aforementioned statistical features could be deceiving sometimes when distributions similar to multi-modal distributions which have similar looks with Fig.1(b) are considered. That's to say, even all the aforementioned indexes are the same for two distributions, the two distributions still look very different. In order to fix this problem, we introduce another index $Beta$ [12]

$$Beta(X) = \frac{Skew(X)^2 + 1}{Kurt(X)} \qquad (14)$$

4.2. Action Classification

Random forest is a powerful classifier belonging to the family of ensemble learning, whose component is decision trees. This ensemble tree bag consists of multiple decision trees. When a classification problem is given to a random forest, each tree might have its own idea and vote, the class that gains most votes would be the result for the final decision.

In order to generate decision trees, if the training dataset has the size of N for each tree, we randomly select N data from the whole dataset and put those N data back to the dataset after usage. This bootstrapping mode is applied for sampling.

Assuming that the dimension of features is M, and we select a constant $m << M$. A subset consisting of m randomly selected element is formed to train each tree. When the tree grows, optimal features are considered. No pruning is required in the structure of random forest.

Each decision tree is a CART (Classification And Regression Tree). The trees are trained by minimizing Gini coefficient in the form of

$$Gini(p) = \Sigma p_k(1 - p_k) \qquad (15)$$

where k is the number of classes, and the probability that a sampled point belongs to class k is p_k.
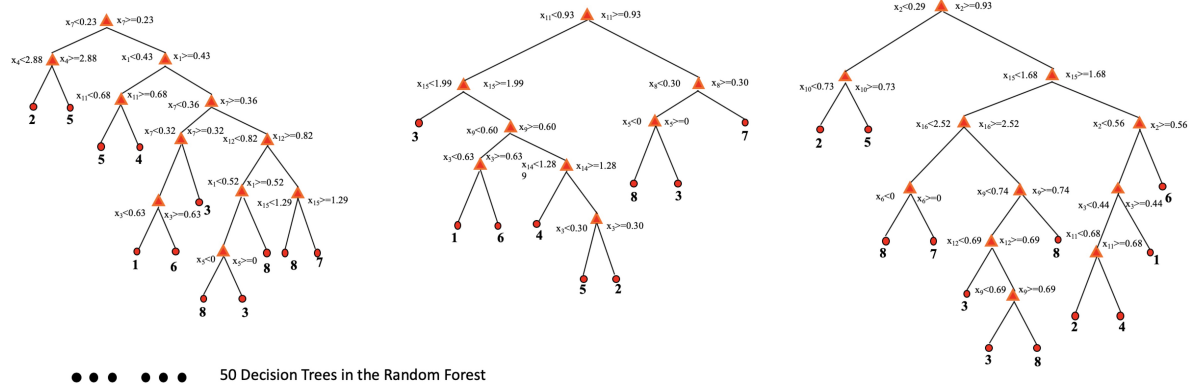
Figure 3. part of trained result of random forest

5. Experimental Results

In order to validate the effectiveness of the proposed algorithm, this section includes four parts. To validate effectiveness of curvature as a signature of action itself, we firstly calculate curvatures on a relatively simpler dataset, which doesn't have much noise or disturbance from background. Features are generated from the curvatures and fed into random forest to obtain the classification result. Robust analysis against flipping is carried out. Then, to further validate our algorithm, Mask R-CNN is customized as a data-augmentation process. And comparison experiment among several different methods are carried out.

5.1. Validation of the Curvature Description

This part aims at validating effectiveness of curvature as a signature of action recognition. The experiments are carried out on Weizmann dataset, whose results are shown in Fig.2. First curvatures (usually refered to as curvature) and second curvatures (usually refered to as torsion) of eight different actions are plotted and analyzed. In Fig2, eight classes of actions are considered, respectively waving with one hand, waving with two hand, jumping side, jumping up, bending, walking, siding, and skipping. The first curvatures are plotted according to equation (4), with respect to universal time scale. It can be observed that the first curvatures of each particular action follow certain particular pattern, and the first curvatures of different actions show different patterns. Second curvatures are calculated according to equation (5).

Theoretically speaking, we could calculate as high order of curvatures as we want. In practice, however, derivative calculation comes with computational error. And as the order of derivative goes higher, this error goes larger. In order to make the calculation of derivative as precise as possible, we used finite difference method to calculate the

Method	Performance
PBMS[36]	0.870
CSTIP[5]	0.967
Vanilla 1st Curvature	0.850
Vanilla 1st, 2nd Curvature	**0.950**

Table 1. Comparison to other methods on on Weizmann dataset.

Method	Performance
MACH[25]	0.692
LTP[38]	0.793
DFCM[21]	0.837
HOG+FV[32]	0.850
Mask R-CNN+1st Curvature	0.759
Mask R-CNN+1st, 2nd Curvature	**0.870**

Table 2. Comparison to other methods on on UCF Sports dataset.

derivatives. And we found out that second curvature, which involves third derivative as the highest order, has acceptable scale of error. And including first curvature and second curvature is enough for the classification task. So there is no need to involve curvatures of higher orders here.

Since the curvatures are obtained, statistical features could be generated from them. Experimental results carried out on Weizmann dataset are shown in TABLE1.

5.2. Robustness Analysis

Another advantage of the proposed curvature signature lies in the fact that it is robust against flipping. This is because the pattern extracted by our algorithm doesn't depend on matching the location of pixels between frames. It could be observed from Fig 5 that symmetric actions result in similar pattern in curvatures, with slight intraclass difference.

Figure 4. Transfer learning result of Mask R-CNN to our task on UCF Dataset [29]

Figure 5. Comparison of curvature when filpping happens on Weizzman Dataset [4]

Thus the curvature as a signature is robust against flipping.

5.3. Data Augmentation

With previous experiments, we proved the effectiveness of curvature as a signature of action. Because background of Weizmann dataset is basically always the same wall not moving, simple matrix subtraction could be carried out to remove the back ground. However, for more complex videos, which are know as videos in the wild, fancier background removal algorithm need to be applied. In order to carry out our algorithm in wild videos, we customize Mask R-CNN[16] as a data augmentation process. Because our algorithm is a few-shot learning, experiments are carried out on a smaller sample of the well-known UCF101 dataset[28], namely UCF Soorts dataset. Fig4 shows the data augmentation result with MASK R-CNN on consecutive frames of a video. It should be noted that MASK R-CNN was trained on COCO dataset which consists of still images. When it comes to videos, motion would introduce extra noises. Background clustering and poor lighting are also potential reasons that could make transfer learning fail. We fine-tuned the last layer of MASK R-CNN in order to accomodate to our task.

5.4. Comparison Experiment

We compare the performances of published methods on Weizmann dataset as well as UCF Sports dataset in TABLE I and TABLE II respectively. It could be observed that first derivative is an effective signature of action, and performance is better when both first and second curvatures are used. We focused most of our narrative on the modeling. It should be noted that the proposed signature of curvature is a universal feature. Classifier is not limited to Random Forest. Other state-of-art classification architectures such as CNN or LSTM could also be incorporated, and multiple features could be incorporated in order to obtain higher precision in classification task.
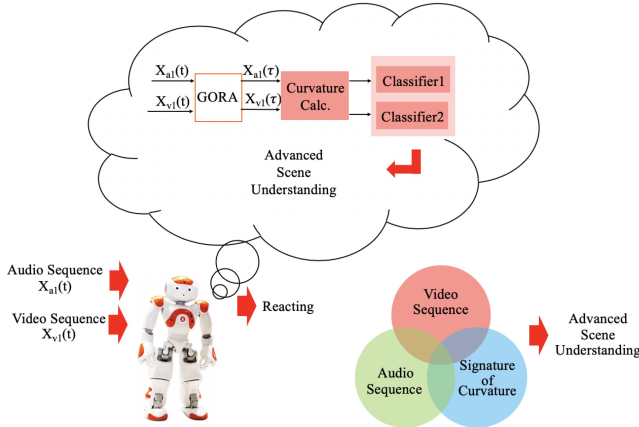
Figure 6. The goal of our future research

6. Conclusions

We proposed a new description of actions based on the curvatures of sequences. This is in contrast to most previous learning-based algorithms, which use edges as description for classification. Building upon the curvature description, features are extracted by calculating statistical indexes. Based on such features, we demonstrate our algorithm in the global time scale instead of a few frames. One limitation of this method is that it is suboptimal for subtle actions with small movement, where noise would do harm to the recognition. Further investigation need to be done in the future about small movements. It should be noted that the proposed method is proved to be effective in but not limited to video sequences. It could also be accommodated for inputs of audio sequence such as language or music. In the future, we would like to look into details of how to customize this algorithm to a fusion of video and audio sequences. As shown in Fig.6, we believe the fusion of video sequence, audio sequence, and signature of curvature would lead to promising performance of advanced scene understanding, which would open a new prospect for smart service robot.

References

[1] A. Alahi, K. Goel, and V. Ramanathan. Social lstm: Human trajectory prediction in crowded spaces. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–971, 2016. 2

[2] J. Ba, K. Swersky, and S. Fidler. Predicting deepzero-shot convolutional neural networks using textual descrip-

tions. *Proceedings of the IEEE international conference on computer vision(ICCV)*, pages 4247–4255, 2015. 1

[3] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould. Dynamic image networks for action recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3034–3042, 2016. 1, 2

[4] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pages 1395–1402. IEEE, 2005. 4, 6

[5] M. Bregonzio, S. Gong, and T. Xiang. Recognising action as clouds of space-time interest points. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1948–1955, 2009. 5

[6] Y. Cao, D. Barrett, and A. Barbu. Recognize human activities from partially observed videos. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2658–2665, 2013. 1

[7] G. Chirikjian. Signal classification in quotient spaces via globally optimal variational calculus. *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 56–64, 2017. 1, 3

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 886–893, 2005. 2

[9] A. Diba, M. Fayyaz, V. Sharmaet, A. Karami, M. Arzani, R. Yousefzadeh, and L. Gool. Temporal 3d convnets using temporal transition layer. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1117–1121, 2018. 1

[10] T. Du, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. *Proceedings of the IEEE international conference on computer vision(ICCV)*, pages 4489–4497, 2015. 2

[11] Y. Du, W. Wang, and L. Wang. Hierarchical recurrent neural network for skeleton based action recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1110–1118, 2015. 2

[12] A. Ellison. Effect of seed dimorphism on the density-dependent dynamics of experimental populations of atriplex triangularis (chenopodiaceae). *American Journal of Botan*, 74(8):1280–1288, 1987. 4

[13] R. Girdhar and D. Ramanan. Attentional pooling for action recognition. *Advances in neural information processing systems (NIPS)*, pages 34–45, 2017. 2

[14] L. Gorman. An analysis of feature detectability from curvature estimation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 235–240, 1988. 2

[15] M. Guo, E. Chou, D. Huang, S. Song, S. Yeung, and F. Li. Neural graph matching networks for few shot 3d action recognition. *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 653–669, 2018. 2

[16] K. He, G. Gkioxari, P. Dollr, and R. Girshick. Mask r-cnn. *Proceedings of the IEEE international conference on computer vision(ICCV)*, pages 2961–2969, 2017. 6

[17] N. Ikizler, R. Cinbis, and P. Duygulu. Human action recognition with line and flow histograms. *International Conference on Pattern Recognition (ICPR)*, pages 1–4, 2008. 2

[18] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2013. 2

[19] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and F. Li. Large-scale video classification with convolutional neural networks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1725–1732, 2014. 2

[20] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *International journal of computer vision*, 1(4):321–331, 1988. 2

[21] T. Lan, Y. Wang, and Greg. Mori. Discriminative figure-centric models for joint action localization and recognition. *Proceedings of the IEEE international conference on computer vision(ICCV)*, pages 2003–2010, 2011. 5

[22] L.Riggs. Curvature as a feature of pattern vision. *Science*, 181(4104):1070–1072, 1973. 2

[23] T. Mitchel, S. Ruan, and G. Chirikjian. Signal alignment for humanoid skeletons via the globally optimal reparameterization algorithm. *IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, pages 217–223, 2018. 1

[24] R. Poppe. A survey on vision-based human action recognition. *Image and vision computing*, 28(6):976–990, 2010. 2

[25] M. Rodriguez. Spatio-temporal maximum average correlation height templates in action recognition and video summarization. 2010. 5

[26] N. Siddharth, A. Barbu, and J. Siskind. Seeing what you're told: Sentence-guided activity recognition in video. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 732–739, 2014. 1

[27] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems (NIPS)*, pages 568–576, 2014. 1

[28] K. Soomro and A. Zamir. Action recognition in realistic sports videos. *Computer vision in sports. Springer, Cham*, pages 181–208, 2014. 6

[29] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 6

[30] R. Sulanke. The fundamental theorem for curves in the n-dimensional euclidean space. 2009. 3

[31] H. Wang, A. Klaser, C. Schmid, and Cheng-Lin Liu. Action recognition by dense trajectories. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3169–3176, 2011. 2

[32] H. Wang and C. Schmid. Action recognition with improved trajectories. *Proceedings of the IEEE international conference on computer vision(ICCV)*, pages 3551–3558, 2013. 2, 5

[33] L. Wang, T. Tan, H. Ning, and W. Hu. Silhouette analysis-based gait recognition for human identification. *IEEE transactions on pattern analysis and machine intelligence*, 25(12):1505–1518, 2003. 1

[34] D. Weinland, R. Ronfard, and E. Boyer. A survey of vision-based methods for action representation, segmentation and recognition.. *Computer vision and image understanding*, 115(2):224–241, 2011. 2

[35] C. Yang, L. Xie, S. Qiao, and A. L. Yuille. Knowledge distillation in generations: More tolerant teachers educate better students. *arXiv preprint arXiv:1805.05551.*, 2018. 2

[36] W. Yang, Y. Wang, and G. Mori. Human action recognition from a single clip per action. *IEEE International Conference on Computer Vision Workshops, ICCV Workshops*, pages 482–489, 2009. 5

[37] L. Yao, A. Torabi, K. Cho, N. Bballas, C. Pal, H. Larochelle, and A. Courville. Describing videos by exploiting temporal structure. *Proceedings of the IEEE international conference on computer vision(ICCV)*, pages 4507–4515, 2015. 1

[38] L. Yeffet and L. Wolf. Local trinary patterns for human action recognition. *Proceedings of the IEEE international conference on computer vision(ICCV)*, pages 492–497, 2009. 5

[39] T. Yu, K. Guo, F. Xu, Y. Dong, Z. Su, J. Zhao, J. Li, Q. Dai, and Y. Liu. Body fusion: Real-time capture of human motion and surface geometry using a single depth camera. *Proceedings of the IEEE international conference on computer vision(ICCV)*, pages 910–919, 2017. 1

[40] B. Zhou, A. Andonian, A. Oliva, and A. Torralba. Temporal relational reasoning in videos. *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 803–818, 2018. 2

[41] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. *Advances in neural information processing systems (NIPS)*, pages 487–495, 2014. 1