2019 8th Brazilian Conference on Intelligent Systems (BRACIS)

Human Action Recognition using 2D Poses

Murilo Varges da Silva*
Federal University of São Carlos - UFSCar
Sao Carlos, SP, Brazil
murilo.varges@ifsp.edu.br

Aparecido Nilceu Marana
Faculty of Sciences - UNESP
Bauru, SP, Brazil
nilceu.marana@unesp.br

***Abstract*—The advances in video capture, storage and sharing technologies have caused a high demand in techniques for automatic recognition of humans actions. Among the main applications, we can highlight surveillance in public places, detection of falls in the elderly, no-checkout-required stores (Amazon Go), self-driving car, inappropriate content posted on the Internet, etc. The automatic recognition of human actions in videos is a challenging task because in order to obtain a good result one has to work with spatial information (e.g., shapes found in a single frame) and temporal information (e.g., movements found across frames). In this work, we present a simple methodology for describing human actions in videos that use extracted data from 2-Dimensional poses. The experimental results show that the proposed technique can encode spatial and temporal information, obtaining competitive accuracy rates compared to state-of-the-art methods.**

***Index Terms*—Human action recognition, Surveillance systems, Spatio-temporal features, Video sequences.**

I. Introduction

Currently, there is a large amount of surveillance cameras installed in several public places (e.g., airports, hospitals, malls, etc.). However, most of these places depend on people working in a monitoring center trying to analyze and detect situations involving human actions and that require attention in real-time. Thus, the need for the development of automatic video understanding techniques for the recognition of human actions in videos has become paramount.

Video understanding is a challenging task and during the last decade increased the interest in this research field. Many researches developed in this area are focused on extracting spatiotemporal features for video understanding. The most relevant methods that deal with hand-crafted feature extraction from videos include those based on spatiotemporal interest points, such as: STIPs (HARRIS3D) [1], SIFT-3D [2], HOG3D [3], MBH [4] and Cuboids [5]. These methods use different encoding schemes based on histograms and pyramids. Another well-known state-of-the-art method is the improved Dense Trajectories (iDT) [6], which presents a good performance in tasks related to video understanding.

Furthermore, there are methods that use body shape analysis, which has been the subject of recent researches for the recognition of human actions in videos [7]–[11]. Those methods usually use silhouette or pose for encoding the human actions. In general, they use the centroids of the silhouettes and their representations are generated by using the distance value from the centroid to each silhouette point, in a radial scheme.

The advantages of shape analysis are the simplicity and the rich information in order to represent human actions.

The disadvantage is that good poses and silhouettes can be difficult to acquire, relying mainly on background subtraction or frame difference, which may fail when parts of the body are occluded.

Due to the development of methodologies that use deep learning in still-image recognition tasks, driven by AlexNet [12], the interest in researches using deep learning techniques applied to videos has increased. Some methods have proposed the use of trained CNNs (Convolution Neural Networks) in images to extract features from individual video frames and then fusing these features into a descriptor with fixed size using pooling and high-dimensional encoding. Other methods use 3D spatiotemporal CNNs, which have been applied in many video understanding problems involving recognition of human actions [13]–[15] and detection of inappropriate content such as pornography [16].

The goal of this paper is to present a light descriptor designed for human action recognition in real-time applications based on differences between angles calculated from the joints of skeletons and their trajectories along the video, computed from 2D poses. The proposed method has achieved state-of-the-art accuracies on two public datasets (KTH [17] and Weizmann [18]), comparable to those obtained by using more sophisticated and expensive techniques.

Besides this introductory section, this paper is organized as follows. Section II presents the proposed human action recognition method. Section III presents the experiments and discusses the results obtained by applying the method on two public datasets. Section IV presents some conclusions of our work.

II. Proposed Method

The proposed method extracts features from 2D human poses obtained from videos by using the OpenPose framework [19]. Figure 1 shows the 25 key points obtained from a 2D pose by such framework.

From the 25 key points obtained from each video frame, 15 (0 to 14) are used in our work to calculate the pose descriptors, which are based on the angles formed by two adjacent straight line segments defined by three key points, and

*The first author is also with the Federal Institute of Education, Science and Technology of São Paulo, Birigui, SP, Brazil.

2643-6264/19/$31.00 ©2019 IEEE
DOI 10.1109/BRACIS.2019.00134

on the trajectories of all key points along L frames. The main steps of our method, proposed to compute the pose descriptors and to classify the pose as a predefined action class, are shown in Figure 2.

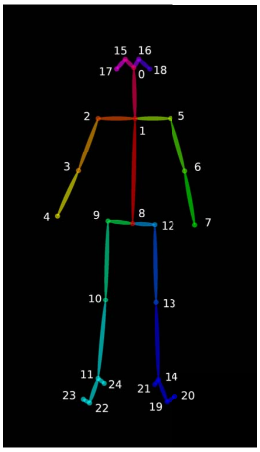


Fig. 1. The skeleton and the 25 key points obtained from a 2D pose by the OpenPose framework [19].

A. Skeleton Angles

In our method we use as descriptors 14 angles calculated from some specific parts of the 2D human skeleton generated by the OpenPose framework, as shown in Table I.

TABLE I
THE 14 ANGLES CALCULATED FROM ADJACENT SKELETON PARTS (PART1 AND PART 2).

Angle	Part 1	Part 2
01	Main Body	Left Shoulder
02	Main Body	Right Shoulder
03	Main Body	Left Hip
04	Main Body	Right Hip
05	Left Shoulder	Neck
06	Right Shoulder	Neck
07	Left Forearm	Left Arm
08	Left Arm	Left Shoulder
09	Right Forearm	Right Arm
10	Right Arm	Right Shoulder
11	Left Thigh	Left Hip
12	Left Thigh	Left Leg
13	Right Thigh	Right Hip
14	Right Thigh	Right Leg

Each skeleton part is represented by a vector $\vec{v}$ defined by two key points $p_i = (x_i, y_i)$ and $p_j = (x_j, y_j)$, according to equation 1:

$$\vec{v} = (x_j - \vec{x_i}, y_j - y_i) \tag{1}$$

Then, the angle θ formed by two adjacent skeleton parts, represented by two vectors, $\vec{v}$ and $\vec{u}$, is calculated according to equation 2:

$$\theta = \arccos\left((\vec{v} * \vec{u})/(||\vec{v}||||\vec{u}||)\right), \tag{2}$$

where $||\vec{v}||$ means the length of vector $\vec{v}$ and the operator $*$ is the dot product of two vectors.

Therefore, for each frame, a feature vector with angles formed by 14 adjacent body parts is generated:

$$Angles = (\theta_1, \theta_2, ..., \theta_{14}) \tag{3}$$

B. Key Points Trajectories

Motivated by Wang et al. [20], who used a descriptor of trajectories for densely sampled points of interest, we used a trajectory descriptor for the key points of the skeleton detected on the video. The structure of the trajectory of one key point P that defines a skeleton part describes the motion pattern of such skeleton part.

Given a trajectory of length L, we encode its shape in a sequence:

$$T = (\Delta P_t, ..., \Delta P_{t+L-1}) \tag{4}$$

of displacement vectors $\Delta P_t = (P_{t+1} - P_t) = (x_{t+1} - x_t, y_{t+1} - y_t)$.

The resulting vector is normalized by the sum of the magnitudes of the displacement vectors:

$$T' = \frac{(\Delta P_t, ..., \Delta P_{t+L-1})}{\sum_{j=t}^{t+L-1} ||\Delta P_j||} \tag{5}$$

As a result, the equations 4 and 5 are repeated for 15 points (0 to 14 Fig. 1) and L frames to form a feature vector of trajectories:

$$Trajectories = (T'_1, T'_2, ..., T'_{15}) \tag{6}$$

C. Feature Encoder Fisher Vector

We use the Fisher Vector (FV) to encode low-level features (Angles and Trajectories) in mid-level features. The FV can be used as a generic framework which combines the benefits of generative and discriminative approaches. In the context of image/video classification, FV has shown to extend the popular Bag-of-Visual-Words (BoVW) by going beyond statistical counting [21].

While BoVW encodes the zero-order statistics of the distribution of descriptors by counting the number of occurrences of visual-codewords, the FV extends the BoVW by encoding the average first and second order differences between the descriptors and visual-codewords.

Fisher Vector [21] encodes both first and second order statistics between the 2D skeleton descriptors and a Gaussian Mixture Model (GMM). We set the number of Gaussians to $K = 20$ and sample all features from the training set to estimate the GMM. Each video is, then, represented by a $K + 2DK$ dimensional Fisher Vector for each descriptor type (Angles and Trajectories), where D is the descriptor dimension, similar to [22]. Finally, we apply power and L2 normalization to the Fisher Vector, as in [21]. Combining the two descriptor types, we concatenate their normalized Fisher Vectors. Finally, a linear SVM is used for classification.

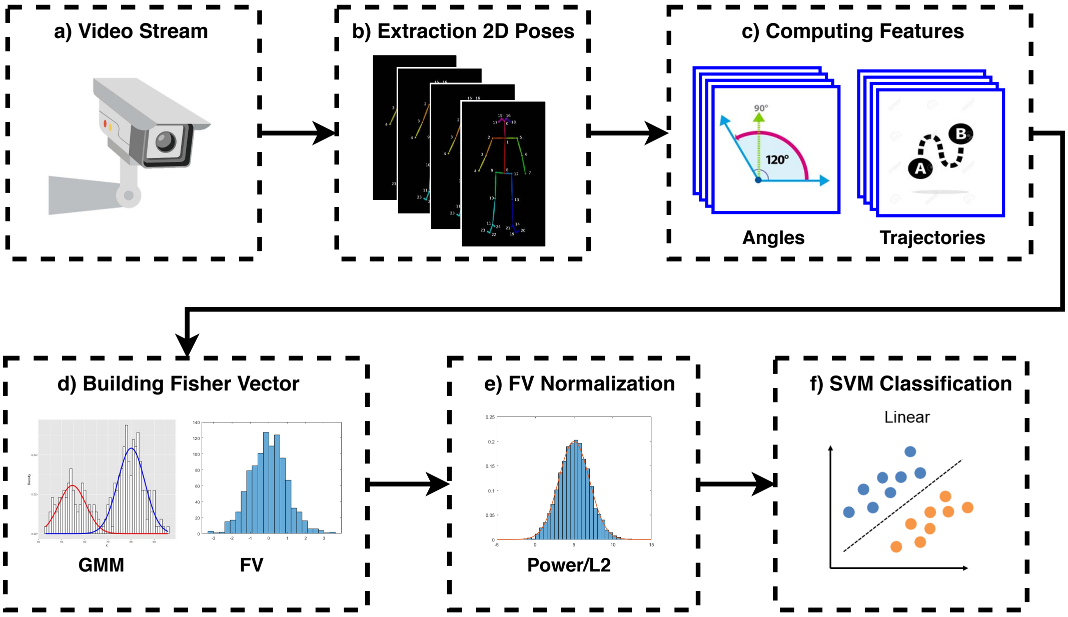


Fig. 2. Illustration of the main steps of our human action recognition method.

III. Experiments and Results

During the experiments an Intel XEON(R) CPU E5620 @2.40GHZ with 16 cores, 39GB of RAM and TITAN XP GPU was used. The 2D pose extraction was performed using OpenPose [19] which was coded in C++ with Caffe framework [23] and used GPU. The feature extraction described in Section II was coded in Python, Fisher Vector encoder and the classification was written in Python using some functions from Scikit-learn [24] without parallel computing. The code for all steps performed is available in GitHub[1].

Two public datasets were used to evaluate our method:

- **KTH [17]:** This dataset contains six classes of human actions (Walking, Jogging, Running, Boxing, Hand waving, Hand clapping) performed several times by 25 people in four different scenarios: outdoors (S1), outdoors with scale variation (S2), outdoors with different clothes (S3) and indoors (S4), as shown in Figure 3. The dataset contains 599 videos acquired in similar backgrounds with a static camera, containing a total of 289,715 frames, 11,375.32 seconds, captured at 25fps and size of 160 X 120 pixels.
- **Weizmann [18]:** This dataset consists of 10 classes (Side, Jack, Bend, Wave1, Wave2, Walk, Skip, Pjump, Jump, Run) of nine actors performing each action, sometimes more than once, resulting in 93 videos. The dataset contains a total of 5,701 frames, 228.04 seconds, captured at 25fps and size of 180 X 144 pixels. All the actions occur on the same static background as shown in Figure 4.

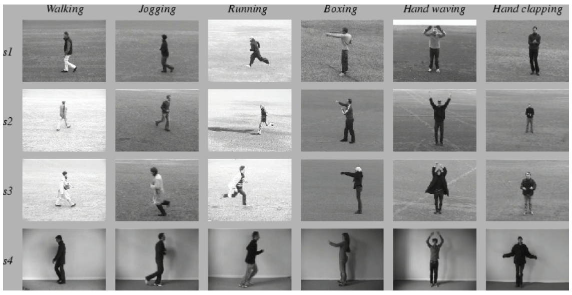


Fig. 3. Sample frames from the KTH [17] dataset for six actions and four scenarios.

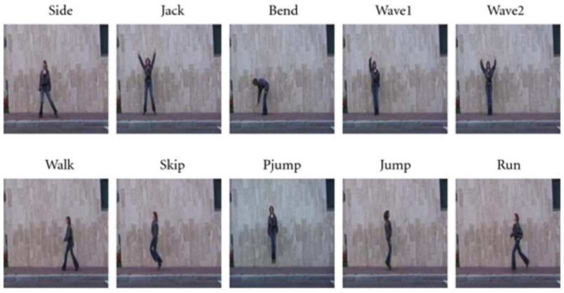
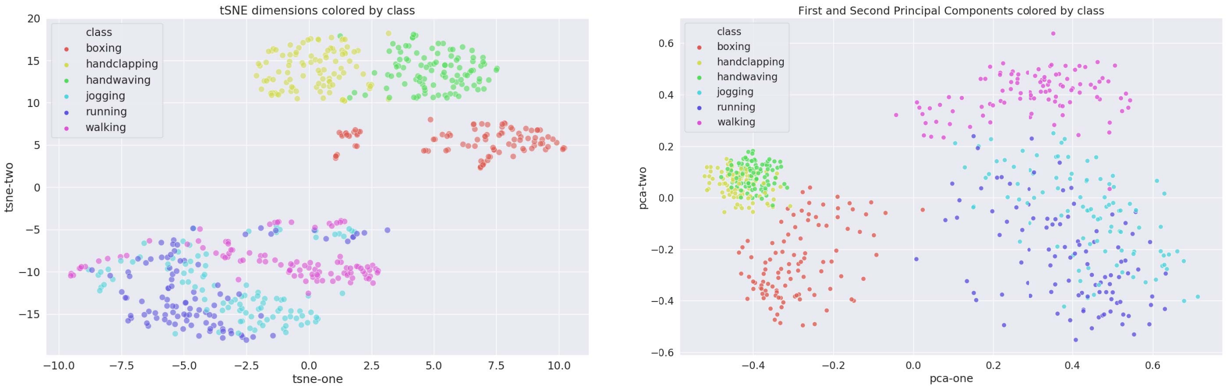


Fig. 4. Sample frames from the Weizmann [18] dataset for ten actions.

A. Features Embedding

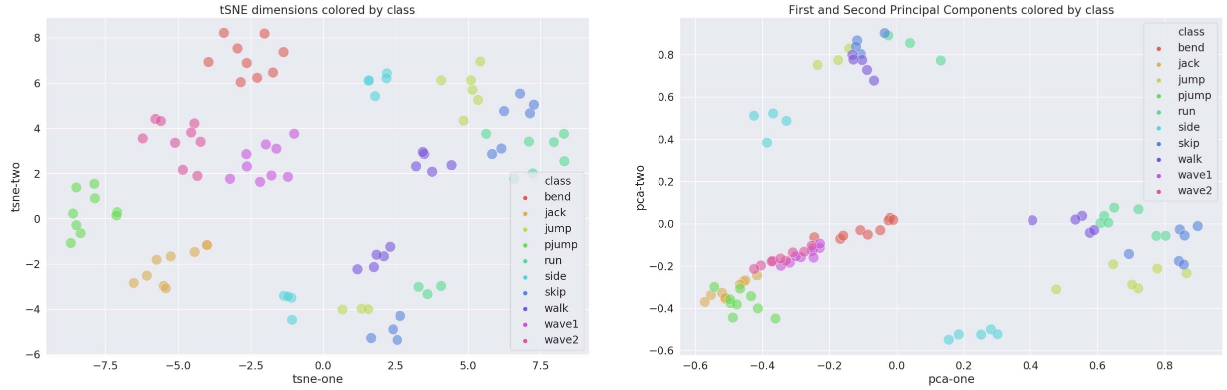
Assessing the representation power of the features extracted from the two datasets (KTH and Weizmann), the features were

[1] https://github.com/murilovarges/HumanActionRecognition2DPoses



(a) (b)

Fig. 5. Feature embedding visualizations of FV on samples from KTH dataset. (a) Using t-SNE and (b) Using PCA. Figure best viewed in color.



(a) (b)

Fig. 6. Feature embedding visualizations of FV on samples from Weizmann dataset. (a) Using t-SNE and (b) Using PCA. Figure best viewed in color.

extracted from all videos and then projected to 2D space using the t-SNE [25] and PCA [26]. Plotting the 2D representation, we use the Fisher Vector obtained from concatenation of Angles and Trajectories as described in Section II-C.

Figure 5 shows the feature embedding from KTH dataset. Figure 5(a) presents the t-SNE and Figure 5(b) presents the PCA. One can see that the features are semantically separable, although some samples of the classes Running and Jogging present some overlapping.

The Weizmann dataset contains fewer samples (93) compared to KTH (599), thus making it easier to classify, as shown in Figure 6. Figure 6(a) presents the t-SNE and Figure 6(b) presents the PCA.

B. Classification

In order to evaluate and compare the proposed method, we used the Leave-One-Out-Cross-Validation (LOOCV) protocol as validation technique. In LOOCV at each iteration, a single sample is taken as the test sequence, while the other $n-1$ samples are used to train the model, which is repeated for all video samples.

To compute the trajectory descriptor (Section II-B), we need to set L (trajectory length) and W (sampling step size) parameters. Based on some experiments, we achieved the best results setting L = 20 and W = 10 (For Weizmann dataset W = 1).

Table II shows the results of the two descriptors presented in this work and their concatenation for the KTH and Weizmann datasets. It is worth noting that only methods that used the LOOCV protocol were presented to avoid the divergences that other protocols can cause in the accuracy of each technique. The results show for both datasets that using the fusion between Angles and Trajectories, we can achieve better results.

TABLE II
ACCURACY RATES (%) FOR KTH AND WEIZMANN DATASETS.

Method	Year	Dataset	
		KTH	Weizmann
FV (Angles + Trajectories)	2019	**95.33**	**97.85**
FV (Angles)	2019	94.32	87.10
FV (Trajectories)	2019	78.96	76.34
Zhang and Tao [27]	2012	93.50	93.87
Junejo and Aghbar [28]	2012	-	88.60
Chaaraoui et al. [29]	2013	-	90.32
Guo et al. [30]	2013	98.50	100
Ravanbakhsh et al. [31]	2015	95.60	-
Doumanoglou et al. [32]	2016	88.70	-
Alcantara et al. [33]	2017	92.20	100
Almeida et al. [34]	2017	96.80	-
Carmona and Climent [35]	2018	97.50	98.80
Chou et al. [36]	2018	90.58	95.56
Singh et al. [37]	2019	94.50	97.66

Figure 7 shown the confusion matrix for the KTH dataset by using the fusion of Angles and Trajectories that achieved 95.33% of accuracy. It is possible to notice that the errors occur mainly between the classes Running and Jogging which have the same spatial pattern and a small temporal difference (speed of the action).

The KTH is a challenging dataset to our method since some classes (Walking, Jogging, Running) present the same spatial pattern, then the approach needs to accurately represent the movement pattern to separate those classes in the classification phase. Despite of it, our method achieved excellent accuracy compared to state-of-the-art methods.

The confusion matrix for the Weizmann dataset is shown in Figure 8. The results presented are for the fusion of Angles and Trajectories that achieved 97.85% of accuracy.

The Weizmann dataset contains small video sequences. Thus, we can use only a few videos for training. However, our method was effective to represent samples and provided excellent classification results compared to the other methods.

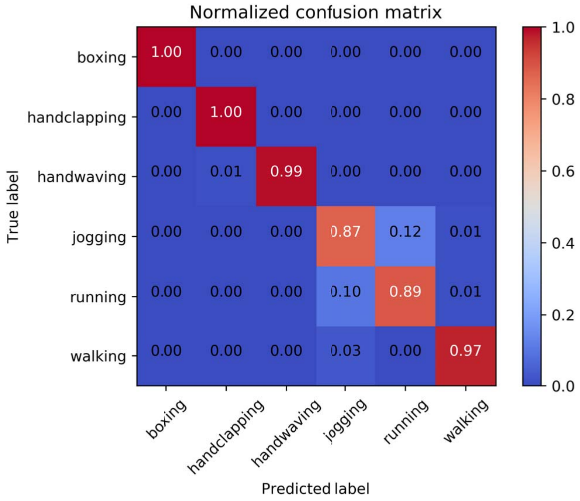


Fig. 7. Confusion Matrix KTH Dataset. Figure best viewed in color.

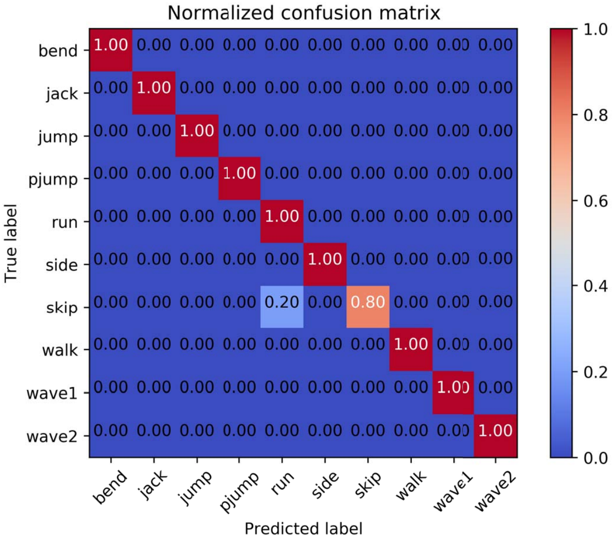


Fig. 8. Confusion Matrix Weizmann Dataset. Figure best viewed in color.

IV. CONCLUSION

Human action recognition in video is a challenging problem. Consequently, the development of a robust method that deals well with any possible action and environment is also a challenge.

This paper presented a new approach to recognize human action in videos by combining information of angles formed by adjacent human skeleton parts and trajectories of skeleton key points (that define skeleton parts) across frames. Our descriptors are easier and lighter to compute comparing to other state-of-the-art methods. Our method is based on 2D poses and there are already several methods that achieve good results in the extraction of 2D poses with real-time processing speed, such as OpenPose [19] and PifPaf [38]. Thus, due to the simplicity of our technique, it can be used in applications that require real-time processing with low computational costs. The results obtained are competitive compared to more sophisticated and complex state-of-the-art methods, like those that rely on dense trajectories and deep learning techniques.

Acknowledgment

We thank NVIDIA Corporation for the donation of the GPU used in this study. This study was financed in part by CAPES - Brazil (Finance Code 001).

References

[1] I. Laptev and T. Lindeberg, "Space-time interest points," in *Proceedings Ninth IEEE International Conference on Computer Vision*, Oct 2003, pp. 432–439 vol.1.

[2] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proceedings of the 15th ACM International Conference on Multimedia*, ser. MM '07. New York, NY, USA: ACM, 2007, pp. 357–360.

[3] A. Klaser, M. Marszalek, and C. Schmid, "A Spatio-Temporal Descriptor Based on 3D-Gradients," in *BMVC 2008 - 19th British Machine Vision Conference*, M. Everingham, C. Needham, and R. Fraile, Eds. Leeds, United Kingdom: British Machine Vision Association, Sep. 2008, pp. 275:1–10.

[4] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *Computer Vision – ECCV 2006*, A. Leonardis, H. Bischof, and A. Pinz, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 428–441.

[5] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, Oct 2005, pp. 65–72.

[6] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *2013 IEEE International Conference on Computer Vision*, Dec 2013, pp. 3551–3558.

[7] M. F. de Alcântara, T. P. Moreira, and H. Pedrini, "Motion silhouette-based real time action recognition," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, J. Ruiz-Shulcloper and G. Sanniti di Baja, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 471–478.

[8] K. Raja, I. Laptev, P. Pérez, and L. Oisel, "Joint pose estimation and action recognition in image graphs," in *2011 18th IEEE International Conference on Image Processing*. IEEE, 2011, pp. 25–28.

[9] A. A. Chaaraoui, P. Climent-Pérez, and F. Flórez-Revuelta, "Silhouette-based human action recognition using sequences of key poses," *Pattern Recognition Letters*, vol. 34, no. 15, pp. 1799–1807, 2013.

[10] S. Singh, S. A. Velastin, and H. Ragheb, "Muhavi: A multicamera human action video dataset for the evaluation of action recognition methods," in *2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*. IEEE, 2010, pp. 48–55.

[11] S. Cheema, A. Eweiwi, C. Thurau, and C. Bauckhage, "Action recognition by learning discriminative key poses," in *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. IEEE, 2011, pp. 1302–1309.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, ser. NIPS'12. USA: Curran Associates Inc., 2012, pp. 1097–1105.

[13] 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*, vol. 35, no. 1, pp. 221–231, Jan 2013.

[14] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *IEEE ICCV*, Washington, DC, USA, 2015, pp. 4489–4497.

[15] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," *CoRR*, vol. abs/1711.11248, 2017.

[16] M. V. da Silva and A. N. Marana, "Spatiotemporal cnns for pornography detection in videos," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, R. Vera-Rodriguez, J. Fierrez, and A. Morales, Eds. Cham: Springer International Publishing, 2019, pp. 547–555.

[17] I. Laptev, B. Caputo *et al.*, "Recognizing human actions: a local SVM approach," in *ICPR*. IEEE, 2004, pp. 32–36.

[18] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2247–2253, December 2007.

[19] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields," in *arXiv preprint arXiv:1812.08008*, 2018.

[20] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action Recognition by Dense Trajectories," in *IEEE Conference on Computer Vision & Pattern Recognition*, Colorado Springs, United States, Jun. 2011, pp. 3169–3176. [Online]. Available: http://hal.inria.fr/inria-00583818/en

[21] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *European conference on computer vision*. Springer, 2010, pp. 143–156.

[22] J. Krapac, J. Verbeek, and F. Jurie, "Modeling spatial layout with fisher vectors for image categorization," in *2011 International Conference on Computer Vision*, Nov 2011, pp. 1487–1494.

[23] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[25] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[26] I. Jolliffe, *Principal component analysis*. Springer, 2011.

[27] Z. Zhang and D. Tao, "Slow feature analysis for human action recognition," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 34, no. 03, pp. 436–450, mar 2012.

[28] I. N. Junejo and Z. Al Aghbari, "Using sax representation for human action recognition," *Journal of Visual Communication and Image Representation*, vol. 23, no. 6, pp. 853–861, 2012.

[29] A. A. Chaaraoui, P. Climent-Pérez, and F. Flórez-Revuelta, "Silhouette-based human action recognition using sequences of key poses," *Pattern Recognition Letters*, vol. 34, no. 15, pp. 1799–1807, 2013.

[30] K. Guo, P. Ishwar, and J. Konrad, "Action recognition from video using feature covariance matrices," *IEEE Transactions on Image Processing*, vol. 22, no. 6, pp. 2479–2494, 2013.

[31] M. Ravanbakhsh, H. Mousavi, M. Rastegari, V. Murino, and L. S. Davis, "Action recognition with image based cnn features," *arXiv preprint arXiv:1512.03980*, 2015.

[32] A. Doumanoglou, N. Vretos, and P. Daras, "Action recognition from videos using sparse trajectories," *IET Conference Proceedings*, pp. 10 (5 .)–10 (5 .)(1), January 2016. [Online]. Available: https://digital-library.theiet.org/content/conferences/10.1049/ic.2016.0078

[33] M. F. de Alcantara, T. P. Moreira, H. Pedrini, and F. Flórez-Revuelta, "Action identification using a descriptor with autonomous fragments in a multilevel prediction scheme," *Signal, image and video processing*, vol. 11, no. 2, pp. 325–332, 2017.

[34] R. Almeida, B. Bustos, Z. K. G. do Patrocínio, and S. J. F. Guimarães, "Human action classification using an extended bow formalism," in *International Conference on Image Analysis and Processing*. Springer, 2017, pp. 185–196.

[35] J. M. Carmona and J. Climent, "Human action recognition by means of subtensor projections and dense trajectories," *Pattern Recognition*, vol. 81, pp. 443 – 455, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320318301493

[36] K.-P. Chou, M. Prasad, D. Wu, N. Sharma, D.-L. Li, Y.-F. Lin, M. Blumenstein, W.-C. Lin, and C.-T. Lin, "Robust feature-based automated multi-view human action recognition system," *IEEE Access*, vol. 6, pp. 15 283–15 296, 2018.

[37] T. Singh and D. K. Vishwakarma, "A hybrid framework for action recognition in low-quality video sequences," *arXiv preprint arXiv:1903.04090*, 2019.

[38] S. Kreiss, L. Bertoni, and A. Alahi, "Pifpaf: Composite fields for human pose estimation," *CoRR*, vol. abs/1903.06593, 2019. [Online]. Available: http://arxiv.org/abs/1903.06593