Human Action Identification in Videos using Descriptor with Autonomous Fragments and Multilevel Prediction

Marlon Fernandes de Alcantara[1], Helio Pedrini
Institute of Computing, University of Campinas
Campinas-SP, Brazil, 13083-852

***Abstract*—Recent technological advances have provided devices with high processing power and storage capacities. Video cameras are found in several places, such as banks, airports, schools, supermarkets, streets, homes and industries. Despite this technological potential, most of the acquired videos are only stored and never analyzed. The flexibility in the use of cameras and computational tools allows their application in areas such as surveillance, strategic planning, crime prevention, manufacturing, traffic monitoring, among others. Video equipments have continuously improved, achieving high resolution rates and frames per second. However, most of the video analysis tasks are still performed by human operators, whose performance may be influenced by factors such stress and fatigue. In order to change such current scenario, this work proposes and evaluates the development of a methodology for identifying common human actions in videos by means of a CMSIP descriptor (Cumulative Motion Shape's Interest Points) applied to a multilevel prediction scheme with retraining. The approach is built by dividing the descriptor into portions that can be considered and interpreted independently by following distinct ways on the classification model, such that, in a later step, a central mechanism will be responsible for deciding which action is being observed in the video sequence. Our method has proved to be fast and with accuracy compatible to the state-of-the-art on known public data sets. Furthermore, the developed prototype demonstrated to be a promising tool for real-time applications.**

***Keywords*-multilevel prediction; action recognition; machine learning; computer vision.**

I. Introduction

Surveillance systems have a wide range of applications and can be used in tasks such as crime prevention, accident monitoring, personal identification, vandalism prevention, among several others [1]. Through the images obtained by video cameras processed by a monitoring system, it is possible to control the activities in complex scenarios and with a large concentration of people, which could be impracticable to be done by a human operator.

The development of digital technology has promoted substantial progress in the area of visual surveillance. Cameras have been developed at higher resolution, smaller dimensions and higher frame rates. Videos acquired by cameras have been recorded in larger quantity due to the increase in storage capacity of the digital media.

[1] Ph.D. Thesis.

In general, current researches focus on the development of intelligent surveillance systems that aim at interpreting human activity, instead of using a passive monitoring system, which is the most commonly employed technology. Intelligent systems may allow the reduction of the necessity of monitoring operators and can help the analysis of images and videos. Nevertheless, intelligent monitoring systems should be capable of automatically extracting complex information of the observed scene and classifying its main events.

The identification of human actions refers directly to the comprehension of human behavior. This understanding involves modeling and classifying actions within a restricted set of rules. The main strategy for this problem is to divide human actions into stages and classify them. The automatic analysis and classification of actions from surveillance cameras can aid or, sometimes, substitute the human monitoring operator. An effective monitoring system can promote the replacement of current passive systems employed in surveillance and improve the identification of events of interest.

This work describes a real-time action identification method based on motion shapes. A new multilevel descriptor is applied multiple times to a single classifier. The algorithm assumes that cumulative motion shapes (CMS) can provide enough information about the action being performed in a video stream. To deal with different possible scenarios of action occurrence, a set of CMS is extracted according to the number of frames. Each CMS is used as an individual entity in the training stage. The proposed action identification method is evaluated on five public datasets (Weizmann, KTH, MuHAVi, IXMAS and URADL).

II. Datasets

There are several public datasets available for action recognition. In several works, the terms *action* and *activity* are used interchangeably. The following is a summary of the datasets used in our experiments. Some samples are shown in Figure 1.

Weizmann [2] consists of 10 classes, with 9 actors performing each action, sometimes with some actors performing them more than once, resulting in 93 videos. The dataset contains a total of $5,701$ frames, 228.04 seconds captured at 25 FPS, size of 180×144 pixels. All the actions occur in the same static background.

(a) Weizmann [2] (b) KTH [3] (c) MuHAVi [4] (d) IXMAS [5] (e) URADL [6]

Fig. 1. Samples extracted from some public datasets.

KTH [3] consists of 6 classes, with 25 actors performing each action, in 4 different scenes, with the exception of one person, that performs one action (hand clapping) in only 3 scenes, resulting in 599 videos. The dataset contains a total of $289,715$ frames, $11,375.32$ seconds captured at 25 FPS, with size of 160×120 pixels. Most videos have camera movement (zooming, panning and tilting).

MuHAVi [4] (Multicamera Human Action Video Data) consists of 17 classes, with 7 actors performing each action, totalling 119 videos. The actions occur in a closed scenario, with 8 cameras surrounding it. The dataset contains a total of $134,085$ frames, $5,368.16$ seconds captured at 25 FPS, size of 720×576 pixels. The MuHAVi dataset has a subset of manually annotated sequences (MuHAVi-MAS), in which the frames are binary images of the silhouette locations. It is divided into 14 primitive actions and it is usually called MuHAVi14 in the literature. This subset, however, has some actions that vary only in direction (for instance, run left and run right) that are rearranged together, forming another subset with 8 classes, called MuHAVi8.

IXMAS (INRIA Xmas Motion Acquisition Sequences) [5] contains 13 classes, however, only 11 are used for validation in the literature. The dataset also offers manually annotated silhouettes. The sequences are recorded in resolution of 390×291 pixels at 23 FPS. The actors choose freely position and orientation to perform the action, where each action is acquired by five cameras in distinct positions (four side and one top view.)

The URADL (University of Rochester Activities of Daily Living) contains 10 activity daily action classes recorded in high resolution (1280×720 pixels) and 30 FPS. The actions are performed by 4 distinct actors in an indoor environment with a fixed camera. The dataset offers short sequences containing only the background to be used in a previous learning for segmentation purpose.

III. Related Work and Contributions

The classifiers are not a novelty in the literature, but their application in action recognition is recent. Typically, they refer to algorithms requiring special cameras, image processing and learning techniques with large amounts of data. This section introduces concepts related to the main stages of an action recognition system and includes a survey of relevant works.

In the computer vision area, the semantic interpretation of information can be obtained through image or video analysis algorithms. In action recognition, this process may include the detection and segmentation of motion. Subsequent steps of the classification process strongly depend on this step.

Tables I, II, III and IV present accuracy results for our method and others available in the literature. In all datasets, our work is superior or among the best accuracy values.

TABLE I
Accuracy results for KTH and Weizmann datasets.

Work	Dataset	
	KTH	**Weizmann**
Bregonzio et al. [7]	93.1	96.6
Ryoo and Aggarwal [8]	93.8	-
Sun et al. [9]	94.0	97.8
Wang et al. [10]	-	93.3
Ta et al. [11]	93.0	94.5
Raja et al. [12]	86.6	-
Hsieh et al. [13]	-	98.3
Cheema et al. [14]	-	91.6
Baccouche et al. [15]	92.2	-
Le et al. [16]	93.9	-
Bregonzio et al. [17]	94.3	96.7
Junejo and Aghbari [18]	-	88.6
Zhang and Tao [19]	93.5	93.9
Onofri and Soda [20]	97.0	-
Chaaraoui et al. [21]	-	90.3
Ji et al. [22]	90.2	-
Guo et al. [23]	98.5	100.0
Moghaddam and Piccardi [24]	-	96.8
Alcantara et al. **[25]**†	-	94.6
Tran et al. [26]‡	87.1	-
Alcantara et al. **[27]**†	90.1	96.8
Cai et al. [28]	-	97.9
Antonucci et al. [29]	72.5	74.7
Guo and Chen [30]	94,7	-
Moayedi et al. [31]	100.0	100.0
Yang and Ma [32]‡	96.0	-
Chen et al. [33]‡	97.1	-
Zhu and Xia [34]	-	98.5
Han and Li [35]	95.2	99.2
Alcantara et al. **[36]**†	89.1	97.4
Alcantara et al. **[37]**†	**92.2**	**100.0**

†Work developed during the Ph.D. thesis.

‡KTH validation using a subdivision into different scenarios.

IV. Methodology

In our methodology, an action is considered as a set of motion patterns developed over time. However, instead of using shapes in a bag-of-words [9], [17], [20], [56], our work

TABLE II
ACCURACY RESULTS FOR MUHAVI, MUHAVI14 AND MUHAVI8 DATASETS.

Work	Dataset		
	MuHAVi	**MuHAVi14**	**MuHAVi8**
Wu and Jia [38]	69.2	-	-
Singh et al. [4]	-	97.8	82.4
Moghaddam and Piccardi [39]	80.4	-	-
Karthikeyan et al. [40]	88.2	-	-
Cheema et al. [14]	-	86.0	95.6
Moghaddam and Piccardi [24]	92.0	-	-
Chaaraoui et al. [21]	-	91.2	97.1
Chaaraoui and Flórez-Revuelta [41]	-	98.5	100.0
Alcantara et al. **[27]**†	89.1	94.1	**100.0**
Cai et al. [28]	-	98.5	-
Alcantara et al. **[36]**†	91.6	**95.6**	**100.0**
Alcantara et al. **[37]**†	**92.4**	**95.6**	**100.0**

†Work developed during the Ph.D. thesis.

TABLE III
ACCURACY RESULTS FOR IXMAS DATASET.

Work	Accuracy (%)
Weinland et al. [42]	57.9
Evgeniou and Pontil [43]	78.2
Farhadi and Tabrizi [44]	58.1
Reddy et al. [45]	72.6
Liu et al. [46]	75.3
Junejo et al. [47]	72.7
Li and Zickler [48]	81.2
Li et al. [49]	90.5
Huang et al. [50]	57.3
Wu and Jia [51]	88.8
Yan et al. [52]	82.5
Alcantara et al. **[36]**	**81.1**

TABLE IV
ACCURACY RESULTS FOR URADL DATASET.

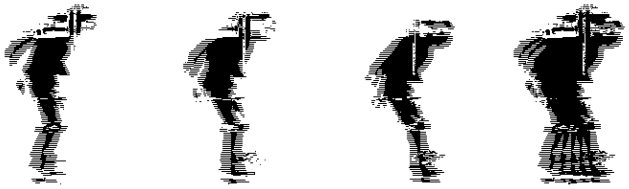
Work	Accuracy (%)
Bobick and Davis [53]†	33.0
Dollar et al. [54]†	36.0
Laptev et al. [55]†	59.0
Messing et al. [6]	63.0
Messing et al. [6]	67.0
Messing et al. [6]	89.0
Alcantara et al. **[36]**	**88.0**

†Experiments performed by Messing et al. [6].

employs every set of indexed points of a silhouette as a fragment that contributes independently to identify the action. Figure 2 shows a diagram with the main stages of our action identification methodology.

To acquire the silhouette of a person performing an action, a motion segmentation is initially performed. This information can be obtained when analyzing adjacent frames, or, for long sequences, the history of all frames until the current instant. The motion segmentation returns body parts or an entire person involved in the action. For instance, in the two-hands wave, only the hands and arms are segmented. In this work, the silhouette is the portion of the body that moves to perform a specific action, therefore, the silhouette is not perfectly extracted.

The number of frames used in CMS composition remains the same for the entire training and prediction steps in a determined dataset. Figure 3 shows a sample of a CMS composition where (a)-(c) three motion shapes can be seen and (d) its union resulting in the built CMS.



(a) Frame 1 (b) Frame 2 (c) Frame 3 (d) CMS

Fig. 3. (a)-(c) Samples of an inaccurate motion extraction in a jump action from Weizmann dataset; (d) resulting CMS from the union of (a) to (c) [57].

The algorithm for interest point detection selects the points of the motion shape that are closest to the control points fixed in a bounding box. These fixed points are called key points and are equally spaced in the bounding box. The number of key points is parameterized, however, it is always constant in a same dataset.

Let c_a, c_b, c_c and c_d be the four corners of a bounding box in clockwise direction. Point p represents the k-th subdivision between two adjacent corners, denoted by p_k in Equation 1, where D is the number of subdivisions in the bounding box, x is any of the corners and y is the subsequent corner in clockwise direction.

$$p_k = \frac{k.(c_x - c_y)}{D} + c_x \qquad k \rightarrow 0 \ldots D-1 \qquad (1)$$

There is no relation of order or priority among CMSs in a video stream. Extracting multiple samples from the same sequence allows for the learning process to be independent of when the action starts. For example, a running person can start the action with both feet together or with one foot ahead.

Finally, a normalization between $[-1, 1]$ is applied to keep the center of the bounding box at the origin of Cartesian plane. This normalization makes the classifier robust to scale variations among distinct video streams or even zoom effects in a same stream. The normalized coordinates of each interest point form the action descriptor.

A. Descriptor Construction

The descriptor is built through the union of normalized coordinates of interest points obtained from a constant number of CMSs. To guarantee that the number of CMSs in the descriptor is constant, a sampling of N points is applied, where N is proportional to the action duration observed in the specific dataset.

The construction of the descriptor considers that many of the keypoints will have little (or no) influence on the classification

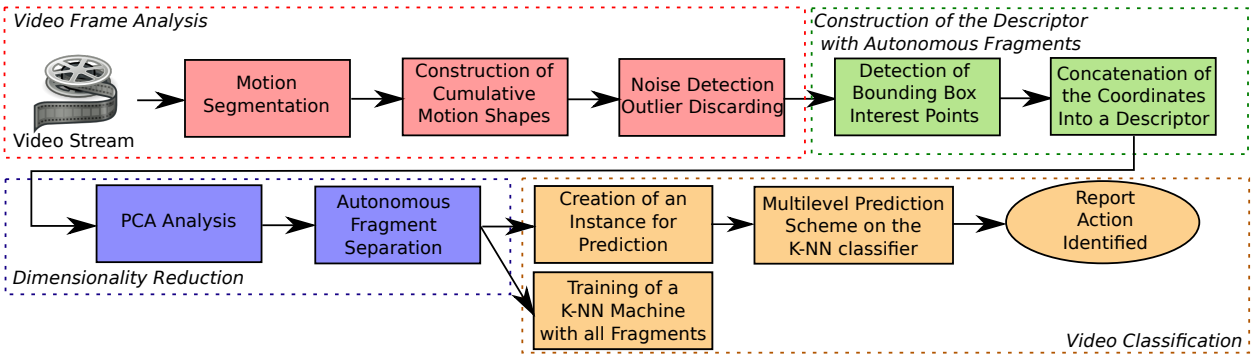


Fig. 2. Main stages of the proposed action identification methodology.

process. Principal Component Analysis [58] is used to reduce the dimensionality of the descriptor, while maintaining only its most important features. This makes the descriptor compact enough to be used by the classifiers. It is worth mentioning that there is no order relation among the samples.

B. Machine Learning

SVM and k-NN classifiers are used in several works for supervised training. We also used both classifiers to validate the proposed method. The training and prediction stages perform multiple operations in the classifier, one for each CMS.

For the SVM, a variation known as one-against-all was used for multiclass purposes. Since k-NN is natively multiclass, to insert a CMSIP into the descriptor is only necessary to train the set of CMSs that belong to the CMSIP in the same classifier.

C. Training and Prediction

Considering a CMSIP formed by a set of CMSs, each CMS is used for training independently and it has a label corresponding to the action associated with it. The training process is done multiple times, where each training step depends on the instances and results obtained from the previous prediction. The classifier chosen for the final stage is k-NN since it proved to be substantially faster. Training time is a critical factor in our method and SVM was not advantageous in terms of accuracy.

The value k for k-NN is an input parameter whose value is proportional to the amount of video sequences and the number of frames in each sequence. This value indicates the number of neighbors that should be considered in the prediction step.

In addition to the value k, a value k' ($k' \leq k$) was considered in our work. Thus, if an item of data to be searched does not contain at least k' votes, it is not considered as a valid vote and it is not included in the final count. This scheme allows for a blank vote when the classifier does not have confidence in the resulting response.

In the prediction step, the classifier identifies the less likely class and remove it from the set of possible classes for the instance being tested. The least likely class is identified as follows: for each descriptor attribute, a prediction is performed and this attribute is classified as one of the possible classes as a vote or it is classified as an outlier, such that its vote is not counted. At the end, after summing up the votes for each class, the class with less votes is discarded. When a class is removed, it is necessary to retrain the k-NN classifier with the remaining classes. After consecutive removals, there will be only a single class and, trivially, it is possible to conclude that this will be the final class to be represented by the descriptor. This response is sent to the output, ending the classification process.

V. Results

As mentioned in the previous section, k-NN and SVM classifiers were chosen for the tests. The parameters used in the descriptor were obtained from exhaustive grid search. They differ according to the databases due to image resolution, amount of motion, among other factors. The thesis presents a deeper discussion on the parameter configuration.

The method adopted for training and testing was the leave-one-out. Although this technique requires intensive processing time, it provides an accurate assessment of the classification results.

The best values found for each parameter are shown in Table V. The following parameter values are reported: number of forms used in the construction of CMSIP (NF), number of CMSIPs to be sampled (NS), dimensions of the array of control points (DX and DY), number of dimensions for PCA algorithm (ND), and values of k and k' for k-NN algorithm. The last column shows the accuracy and execution time, respectively.

TABLE V
Parameters employed in our experiments to achieve high accuracy.

Dataset	NF	NS	DX, DY	ND	k, k'	Accuracy (%)
Weizmann	2	40	8, 4	34	2.2	100.0
KTH	4	30	8, 8	18	6.2	92.2
MuHAVi	6	80	20, 10	55	8.3	92.4
MuHAVi14	4	40	16, 8	35	4.2	95.6
MuHAVi8	2	20	16, 8	32	2.1	100.0
IXMAS	2	50	8, 8	30	1.1	82.6
URADL	2	25	16, 16	60	2.2	90.7

Despite the multilevel prediction process, the algorithm is efficient enough to run in real-time applications. As shown in Table VI, the number of frames per second (FPD) achieved by the proposed method is higher than 24 for the majority of the evaluated datasets, which is typically required in surveillance applications. The exception was the URADL, which was processed at 10.24 frames per second. It is important to mention that this specific dataset contains high-resolution video sequences, which is not usually common in surveillance.

TABLE VI
TIME REQUIRED IN THE FEATURE EXTRACTION AND CLASSIFICATION PROCESSES.

Datasets	Extraction (s)	Classification (s)	Frames	FPS
Weizmann	$4,85$	$0,270$	5.701	$1113,48$
KTH	$1.347,38$	$5,382$	289.715	$214,17$
MuHAVi	$2.850,29$	$1,504$	137.085	$48,01$
IXMAS	$865,464$	$2,308$	34.155	$39,36$
URADL	$7.100,6$	$4,732$	72.729	$10,24$

VI. PUBLICATIONS

The following papers have been published during the development of this Ph.D. thesis. The first paper was published in 2013, only one year after the beginning of the research. A journal paper has been recently accepted for publication.

1) *Action Identification using a Descriptor with Autonomous Fragments in a Multilevel Prediction Scheme*. Signal, Image and Video Processing (Accepted for Publication). CAPES/QUALIS A2.
2) *Real-Time Action Recognition Using a Multilayer Descriptor with Variable Size*. Journal of Electronic Imaging (JEI), vol. 25, n. 1, pp. 013020.1-013020.9, February 2016. CAPES/QUALIS A2.
3) *Fast and Accurate Gesture Recognition Based on Motion Shapes*. In: 20th Iberoamerican Congress on Pattern Recognition (CIARP), 2015, Montevideo, Uruguay. Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, 2015. v. 9423. p. 247-254. CAPES/QUALIS B2.
4) *Real-Time Action Recognition Based On Cumulative Motion Shapes*. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP), June 2014, Florence - Italy. v. 1. p. 2917-2921. CAPES/QUALIS A1.
5) *Motion Silhouette-Based Real Time Action Recognition*. In: 18th Iberoamerican Congress on Pattern Recognition (CIARP), November 2013, Havana, Cuba. Lecture Notes in Computer Science, 2013. v. 8259. p. 471-478. CAPES/QUALIS B2.

VII. CONCLUSIONS

This work presented a solution for automatic identification of human actions in videos based on motion shape descriptors using a multilevel prediction scheme with retraining.

In addition to the publications, this study contributed to the construction of a new light descriptor based on Cumulative Motion Shape's Interest Points (CMSIP) that can be used in any video sequence.

The work proposes a novel training and prediction methodology using an algorithm based on fragmented descriptors, the Multilevel Prediction Scheme (MPS), to converge the multiple classification responses and increase the accuracy.

The method was applied to several databases with multiple actions and different setups for camera and environment. The research provides new insights into the action identification problem, providing a solution that can be expanded to other scenarios and more complex actions.

ACKNOWLEDGMENTS

The authors are thankful to FAPESP (grant #2012/20738-1 and #2011/22749-8) for the financial support.

REFERENCES

[1] W. Hu, T. Tan, L. Wang, and S. Maybank, "A Survey on Visual Surveillance of Object Motion and Behaviors," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 34, no. 3, pp. 334–352, Aug. 2004.

[2] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as Space-Time Shapes," in *The Tenth IEEE International Conference on Computer Vision*, 2005, pp. 1395–1402.

[3] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing Human Actions: A Local SVM Approach," in *17th International Conference on Pattern Recognition*, vol. 3, Aug. 2004, pp. 32–36.

[4] S. Singh, S. A. Velastin, and H. Ragheb, "MuHAVi: A Multicamera Human Action Video Dataset for the Evaluation of Action Recognition Methods," in *Advanced Video and Signal Based Surveillance*, 2010, pp. 48–55.

[5] D. Weinland, R. Ronfard, and E. Boyer, "Free Viewpoint Action Recognition using Motion History Volumes," *Computer Vision and Image Understanding*, vol. 104, no. 2-3, pp. 249–257, 2006.

[6] R. Messing, C. Pal, and H. Kautz, "Activity Recognition Using the Velocity Histories of Tracked Keypoints," in *IEEE International Conference on Computer Vision*. Washington, DC, USA: IEEE Computer Society, 2009.

[7] M. Bregonzio, S. Gong, and T. Xiang, "Recognising Action as Clouds of Space-Time Interest Points," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 1948–1955.

[8] M. S. Ryoo and J. K. Aggarwal, "Semantic Representation and Recognition of Continued and Recursive Human Activities," *International Journal of Computer Vision*, vol. 82, no. 1, pp. 1–24, Apr. 2009.

[9] X. Sun, M. Chen, and A. Hauptmann, "Action Recognition via Local Descriptors and Holistic Features," in *Computer Vision and Pattern Recognition*, Miami, FL, USA, 2009, pp. 58–65.

[10] S. Wang, K. Huang, and T. Tan, "A Compact Optical Flowbased Motion Representation for Real-Time Action Recognition in Surveillance Scenes," in *International Conference on Image Processing*, Cairo, Egypt, Nov. 2009, pp. 1121–1124.

[11] A. P. Ta, C. Wolf, G. Lavou, A. Baskurt, and J.-M. Jolion, "Pairwise Features for Human Action Recognition," in *International Conference on Pattern Recognition*, Istanbul, Turkey, Aug. 2010, pp. 3224–3227.

[12] K. Raja, I. Laptev, P. Perez, and L. Oisel, "Joint Pose Estimation and Action Recognition in Image Graphs," in *IEEE International Conference on Image Processing*, Brussels, Belgium, Sep. 2011, pp. 25–28.

[13] C. Hsieh, P. Huang, and M. Tang, "The Recognition of Human Action Using Silhouette Histogram," in *Australasian Computer Science Conference*, M. Reynolds, Ed., vol. 113. Perth, Australia: ACS, 2011, pp. 11–16.

[14] S. Cheema, A. Eweiwi, C. Thurau, and C. Bauckhage, "Action Recognition by Learning Discriminative Key Poses," in *International Conference on Computer Vision*, Barcelona, Spain, 2011, pp. 1302–1309.

[15] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Sequential Deep Learning for Human Action Recognition," in *Human Behavior Understanding*. Springer, 2011, pp. 29–39.

[16] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning Hierarchical Invariant Spatio-Temporal Features for Action Recognition with Independent Subspace Analysis," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2011, pp. 3361–3368.

[17] M. Bregonzio, T. Xiang, and S. Gong, "Fusing Appearance and Distribution Information of Interest Points for Action Recognition," *Pattern Recognition*, vol. 45, no. 3, pp. 1220–1234, Mar. 2012.

[18] I. N. Junejo and Z. A. Aghbari, "Using SAX Representation for Human Action Recognition," *Journal of Visual Communication and Image Representation*, vol. 23, no. 6, pp. 853–861, Aug. 2012.

[19] Z. Zhang and D. Tao, "Slow Feature Analysis for Human Action Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 3, pp. 436–450, 2012.

[20] L. Onofri and P. Soda, "Combining Video Subsequences for Human Action Recognition," in *International Conference on Pattern Recognition*, Tsukuba, Japan, 2012, pp. 597–600.

[21] 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, smart Approaches for Human Action Recognition.

[22] 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, 2013.

[23] K. Guo, P. Ishwar, and J. Konrad, "Action Recognition From Video Using Feature Covariance Matrices," *Image Processing*, vol. 22, no. 6, pp. 2479–2494, 2013.

[24] Z. Moghaddam and M. Piccardi, "Training Initialization of Hidden Markov Models in Human Action Recognition," *Automation Science and Engineering*, vol. 36, no. 99, pp. 1–15, 2013.

[25] M. Alcantara, T. Moreira, and H. Pedrini, "Motion Silhouette-Based Real Time Action Recognition," in *18th Iberoamerican Congress on Pattern Recognition*, vol. 8259, Lecture Notes in Computer Science, Havana, Cuba, Nov. 2013, pp. 471–478.

[26] A. Tran, J. Guan, T. Pilantanakitti, and P. Cohen, "Action Recognition in the Frequency Domain," *arXiv preprint arXiv:1409.0908*, 2014.

[27] M. Alcantara, T. Moreira, and H. Pedrini, "Real-Time Action Recognition Based on Cumulative Motion Shapes," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. Florence, Italy: IEEE, May 2014, pp. 2917–2921.

[28] J. Cai, X. Tang, and G. Feng, "Learning Pose Dictionary for Human Action Recognition," in *International Conference on Pattern Recognition*, vol. 1, Aug 2014, pp. 381–386.

[29] A. Antonucci, R. D. Rosa, A. Giusti, and F. Cuzzolin, "Robust Classification of Multivariate Time Series by Imprecise Hidden Markov Models," *International Journal of Approximate Reasoning*, vol. 56, no. Part B, pp. 249–263, 2015.

[30] W. Guo and G. Chen, "Human Action Recognition via Multi-Task Learning Base on Spatial-Temporal Feature," *Information Sciences*, vol. 320, pp. 418–428, 2015.

[31] F. Moayedi, Z. Azimifar, and R. Boostani, "Structured Sparse Representation for Human Action Recognition," *Neurocomputing*, vol. 161, pp. 38–46, 2015.

[32] J. Yang and Z. Ma, "Action Recognition Using Polyhedron Neighborhood Features," *International Journal of Multimedia and Ubiquitous Engineering*, vol. 10, no. 1, pp. 391–402, 2015.

[33] M. M. Chen, L. Gong, T. Wang, and Q. Feng, "Action Recognition using Lie Algebrized Gaussians over Dense Local Spatio-Temporal Features," *Multimedia Tools and Applications*, vol. 74, no. 6, pp. 2127–2142, 2015.

[34] S. Zhu and L. Xia, "Human Action Recognition Based on Fusion Features Extraction of Adaptive Background Subtraction and Optical Flow Model," *Mathematical Problems in Engineering - Hindawi Publishing Corporation*, vol. 387464, 2015.

[35] H. Han and X. Li, "Human Action Recognition with Sparse Geometric Features," *The Imaging Science Journal*, vol. 63, no. 1, pp. 45–53, 2015.

[36] M. Alcantara, T. Moreira, and H. Pedrini, "Real-time Action Recognition using a Multilayer Descriptor with Variable Size," *Journal of Electronic Imaging*, vol. 25, no. 1, pp. 013 020–013 020, 2016a.

[37] M. Alcantara, T. 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 (Accepted for Publications)*, 2016b.

[38] C. Wu, A. H. Khalili, and H. Aghajan, "Multiview Activity Recognition in Smart Homes with Spatio-Temporal Features," in *International Conference on Distributed Smart Cameras*, Atlanta, GA, USA, 2010, pp. 142–149.

[39] Z. Moghaddam and M. Piccardi, "Histogram-Based Training Initialisation of Hidden Markov Models for Human Action Recognition," in *International Conference on Advanced Video and Signal Based Surveillance*, Boston, MA, USA, 2010, pp. 256–261.

[40] S. Karthikeyan, U. Gaur, B. S. Manjunath, and S. Grafton, "Probabilistic Subspace-based Learning of Shape Dynamics Modes for Multi-view Action Recognition," in *International Conference on Computer Vision*, Barcelona, Spain, 2011, pp. 1282–1286.

[41] A. Chaaraoui and F. Flórez-Revuelta, "Human Action Recognition Optimization based on Evolutionary Feature Subset Selection," in *Genetic and Evolutionary Computation Conference*, New York, NY, USA, 2013, pp. 1229–1236.

[42] D. Weinland, E. Boyer, and R. Ronfard, "Action Recognition from Arbitrary Views using 3D Exemplars," in *IEEE 11th International Conference on Computer Vision*, Oct. 2007, pp. 1–7.

[43] A. Evgeniou and M. Pontil, "Multi-task Feature Learning," *Advances in Neural Information Processing Systems*, vol. 19, p. 41, 2007.

[44] A. Farhadi and M. Tabrizi, "Learning to Recognize Activities from the Wrong View Point," in *European Conference on Computer Vision*, ser. Lecture Notes in Computer Science, D. Forsyth, P. Torr, and A. Zisserman, Eds. Springer Berlin Heidelberg, 2008, vol. 5302, pp. 154–166.

[45] K. Reddy, J. Liu, and M. Shah, "Incremental Action Recognition using Feature-Tree," in *IEEE 12th International Conference on Computer Vision*, Sep. 2009, pp. 1010–1017.

[46] J. Liu, M. Shah, B. Kuipers, and S. Savarese, "Cross-View Action Recognition via View Knowledge Transfer," in *IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2011, pp. 3209–3216.

[47] I. Junejo, E. Dexter, I. Laptev, and P. Perez, "View-Independent Action Recognition from Temporal Self-Similarities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 1, pp. 172–185, Jan. 2011.

[48] R. Li and T. Zickler, "Discriminative Virtual Views for Cross-View Action Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2012, pp. 2855–2862.

[49] B. Li, O. Camps, and M. Sznaier, "Cross-View Activity Recognition using Hankelets," in *IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2012, pp. 1362–1369.

[50] C.-H. Huang, Y.-R. Yeh, and Y.-C. Wang, "Recognizing Actions across Cameras by Exploring the Correlated Subspace," in *European Conference on Computer Vision - Workshops and Demonstrations*, ser. Lecture Notes in Computer Science, A. Fusiello, V. Murino, and R. Cucchiara, Eds. Springer Berlin Heidelberg, 2012, vol. 7583, pp. 342–351.

[51] X. Wu and Y. Jia, "View-Invariant Action Recognition Using Latent Kernelized Structural SVM," in *European Conference on Computer Vision*, ser. Lecture Notes in Computer Science, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds. Springer Berlin Heidelberg, 2012, vol. 7576, pp. 411–424.

[52] Y. Yan, G. Liu, E. Ricci, and N. Sebe, "Multi-Task Linear Discriminant Analysis for Multi-View Action Recognition," in *20th IEEE International Conference on Image Processing*, Sep. 2013, pp. 2842–2846.

[53] A. Bobick and J. Davis, "The Recognition of Human Movement using Temporal Templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257–267, Mar. 2001.

[54] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior Recognition via Sparse Spatio-Temporal Features," in *International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 2005, pp. 65–72.

[55] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning Realistic Human Actions from Movies," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[56] M. Chen and A. Hauptmann, "MoSIFT: Recognizing Human Actions in Surveillance Videos," Carnegie Mellon University Computer Science, Forbes Avenue Pittsburgh, PA, USA, Tech. Rep., 2009.

[57] M. Alcantara, "Identificação de Ações Humanas em Vídeos Utilizando Descritor de Fragmentos Autônomos e Predição Multinível," Tese de Doutorado em Ciência da Computação, Universidade Estadual de Campinas, Campinas/SP, 2015.

[58] I. Jolliffe, *Principal Component Analysis*, 2nd ed. Springer, 2002.