The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-2/W9-2025
ISPRS Intl. Workshop “Photogrammetric and computer vision techniques for environmental and infraStructure monitoring, Biometrics and Biomedicine”
PSBB25 , 9–11 June 2025, Moscow, Russia

Human Action Recognition from Motion Capture Data based on Curve Matching

Evgeny Myasnikov [1]

[1] Samara National Research University, Samara, Russia - mevg@geosamara.ru

Keywords: Action Recognition, Human Actions, Motion Capture Data, Frechet distance, Dynamic Time Warping, Berkeley Multimodal Human Action Database.

Abstract

Human action recognition can be used in a wide variety of scenarios in many areas of human activity, such as medicine, public safety, gaming and entertainment, etc. In this paper, we focus on the problem of human action recognition based on data obtained using motion capture systems. To solve this problem, we use an approach based on the transition of original motion capture data to sequences of points in a lower-dimensional subspace and subsequent classification of human actions by matching the trajectories of points in the specified subspace. In particular, to match the trajectories, we explore the Frechet distance and the Dynamic Time Warping distance in both dependent and independent forms. To form a lower-dimensional subspace, we consider two well-proven approaches: the Principal Component Analysis technique and supervised feature selection procedure. We compare obtained results with alternative techniques using open Berkeley Human Action Database.

1. Introduction

Human action recognition can be used in a wide variety of scenarios in many areas of human activity, such as medicine, public safety, gaming and entertainment, etc.

There are many approaches to human action analysis described in the literature (Kong and Fu, 2022), among which are approaches based on silhouette images (Bobick and Davis, 2001; Ahad, 2012) (both silhouette sequences and integral images), approaches based on bag-of-features/words (O’Hara et al., 2011; Ofli et al., 2013), approaches based on common descriptors such as SIFT, HOG, SURF, etc. (Dhulavvagol and Kundur, 2018; Qazi et al., 2017; Noguchi and Yanai, 2012), and their generalizations (Scovanner et al., 2007; Kl¨aser et al., 2008; Willems et al., 2008), approaches based on spatio-temporal motion trajectories extracted using optical flow analysis (Wang et al., 2013; Wang and Schmid, 2013), approaches based on syntactic recognition methods, namely, finite automata and context-free grammars (Pirsiavash and Ramanan, 2014; Ryoo and Aggarwal, 2006), approaches based on hidden Markov models and conditional random fields (Antonucci et al., 2011; Mavroudi et al., 2018), approaches based on human body (skeleton) models, as well as neural network approaches. The latter group includes deep convolutional (Simonyan and Zisserman, 2014; Khan et al., 2024), recurrent networks (Du et al., 2015; Srivastava et al., 2015), etc.

Depending on the type of input data, human action recognition can be performed based on video data obtained from one or more video cameras; data obtained using stereo cameras; data obtained using scene depth sensors; data obtained using accelerometers; and data obtained using motion capture systems (positional tracking) of one type or another.

In this paper, we focus on solving the problem of action recognition based on data obtained using motion capture systems. In this case, recognition is performed only based on data on the spatial position of sensors attached to the monitored areas of the human body. Additional data (context), which can be obtained using sensors of a different type, such as video cameras, are not used.

To solve the problem of human action recognition, we use an approach based on the transition of original motion capture data to a lower-dimensional subspace describing human poses and classify actions by matching the trajectories of points representing human motion in the specified subspace. In particular, to match the trajectories, we explore two known distances, namely, the Frechet distance and the Dynamic Time Warping (DTW) distance. The latter distance is considered in both dependent and independent forms. To form a lower-dimensional subspace, we consider two well-proven approaches: the principal component analysis (PCA) technique and a supervised feature selection procedure. We compare obtained results with the alternative techniques based on bag-of-features and subsequences classification using classical machine-learning techniques. All numerical experiments were performed using the open Berkeley Human Action database.

The paper is structured as follows. The second section is devoted to the description of the baseline approach adopted in this paper for human action recognition. Section 3 describes the data set and numerical experiments that allow us to determine the parameters of the baseline method and evaluate its quality. Section 4 is devoted to the procedure of searching for a reduced space using supervised feature selection procedure, which allows improving the quality of action recognition. Section 5 describes alternative approaches and compares the quality of different methods. The paper ends up with the conclusion and the list of references.

2. Baseline Approach

A person can be considered as a system of rigid segments (bones) connected to each other by movable hinges (joints). Then human motion can be represented as a change in the position of segments over time. The motion capture system provides a description of human motion that is very close to such a model. It shows how the position of sensors attached to the human body parts of interest changes over time.

In this paper, we proceed from the fact that the amount of data sufficient for motion recognition lies on some manifold in the space formed by the original data. Then the motion can be described in a space of lower dimensionality without loss of the

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-2-W9-2025-201-2025 | © Author(s) 2025. CC BY 4.0 License.

quality of the solution of applied problems. This assumption is supported by the fact that many configurations of the model are unlikely, based on the context of the problems being solved, or are unattainable due to the physiological characteristics of a person. It can be assumed that the dimensionality required to solve applied problems can be significantly lower than the original dimensionality of the data describing human motion.

In this paper, to construct compact descriptions of human motion, we try to describe human poses as points in a parameter space of reduced dimensionality. Human motion forms some trajectories in the specified parameter space. Thus, the task of motion analysis is reduced to the analysis of such trajectories.

The recognition method described below takes into account sequences of vector sets describing the position in 3D space of sensors attached to the human body. Each such set describes the position and pose of a person, and the sequence as a whole describes the action (movement) performed by a person. The result of the method is a class label denoting a human action.

The method proposed in this paper consists of several steps. The first step involves preprocessing, which allows transforming a set of 3D coordinates describing the position and pose of a person into a feature vector invariant to position and size.

In the second step, having a set of vectors describing individual poses of different people, we find a subspace of a given dimension that best preserves information about human poses. For this, we use the most commonly used linear dimensionality reduction method - the principal component analysis technique. Then, each movement is represented as a trajectory (sequence of points) in the found space.

In the third step, we classify the new test trajectory by matching it to the known trajectories in the training set. Since a trajectory can be viewed as a curve in M-dimensional space, as well as a time process, we chose two commonly used distances to compare curves: the Frechet distance (Eiter and Mannila, 1994) and the Dynamic Time Warping (DTW) distance (Senin, 2008).

The Frechet distance can be defined as the minimum length of a segment required to connect corresponding points of two curves when moving along these curves in one direction. The discrete formulation of the Frechet distance between two polygonal curves x_1 and x_2 is as follows [adapted from (Aronov et al., 2006)]:

$$\delta(x_1, x_2) = \min_{S} \max_{(t,q)\in S} \left\| x_1^{[t]} - x_2^{[q]} \right\|.$$

Here we consider curves x_1 and x_2 as time sequences $x_1 = x_1^{[1]}, x_1^{[2]}, \ldots x_1^{[T_1]}$ and $x_2 = x_2^{[1]}, x_2^{[2]}, \ldots x_2^{[T_2]}$, and S is a set containing the pairs of discrete points of time (t, q), which satisfy the following conditions:
- order preservation: if $(t, q)\in S$, then $(t-i, q+j)\notin S$, $(t+i, q-j)\notin S$ for any $i, j>0$;
- completeness: for any $t\in\{1,\ldots T_1\}$ there exists $(t, q)\in S$, as well as for any $q\in\{1,\ldots T_2\}$ there exists $(t, q)\in S$.

DTW (Senin, 2008) allows two time sequences to be compared in such a way that changes in the speed of processes in these sequences (acceleration or deceleration) do not affect the final distance estimate. In the basic version of DTW, the distance (dtw) is calculated using the accumulated cost matrix $W = (w_{i,j})$, formed for two sequences x_1 and x_2 using dynamic programming:

$$w_{i,j} = \left\| x_1^{[i]} - x_2^{[j]} \right\|^2 + \min\{w_{i-1,j}, w_{i,j-1}, w_{i-1,j-1}\}.$$

$$\mathrm{dtw}(x_1, x_2) = \sqrt{w_{T_1,T_2}}\ .$$

To calculate DTW distances in multidimensional spaces, both the multidimensional (dependent) DTW and the independent DTW (dtwi) based on a combination of one-dimensional DTW distances are used:

$$\mathrm{dtwi}(x_1, x_2) = \sum_k \mathrm{dtw}(x_{1,k}, x_{2,k}).$$

Here $x_{i,k} = x_{i,k}^{[1]}, x_{i,k}^{[2]}, \ldots x_{i,k}^{[T_i]}$, k=1..M are partial sequences of k-th coordinates (features) of the original sequences x_i.

To perform the actual classification of actions, a nearest neighbor (NN) classifier is used, based on the above distances.

3. Dataset and Experiments

For the experimental study, we use the open dataset Berkeley Multimodal Human Action Database (Ofli et al., 2013). This dataset contains video sequences taken from different angles, as well as data on human movement of other modalities. In the described study, only data from this dataset obtained by the Impulse motion capture system manufactured by PhaseSpace Inc. were used.

The data represent pre-processed movement trajectories of 43 sensors attached to the human body and recording their position in 3D space at a frequency of 480 Hz. The dataset contains trajectories for 12 people performing 11 different actions. Each action is performed five times, which gives a total of 660 sequences. In accordance with (Ofli et al., 2013), the division into training and test sets was performed by individuals performing the actions. The training set included sequences of seven different people, and the test set included sequences of other five people.

The overall classification accuracy was used as a quality indicator, defined as the proportion of correctly recognized actions out of the total number of test actions.

At the preliminary stage of the research, we determined suboptimal parameters of the method, such as the time step of the movement trajectory and the dimension of the pose subspace formed by the principal component analysis technique. Some results of the study including the dependence of the recognition quality on the time step and on the dimensionality of the pose descriptions for low dimensions are shown in Figure 1.

As expected for both considered distances, the classification quality a gradually decreases with increasing time step s. At the same time, the Frechet distance turns out to be less sensitive to the time step up to s=60. The dependence of the classification quality a on the pose space dimension m for both distances has a local maximum in the region of small dimensions m=2...7.

The best achieved value for the Frechet distance with pose space dimension m=4 and time step s=2 is a=0.796. The best value for the multidimensional (dependent) DTW distance turns out to be a=0.916 with the dimension m=5. At the same time,

the use of independent DTW distance (dtwi) allows to obtain advantages at higher dimensions (see Fig. 2). Thus, at a dimension of m=48, the accuracy is a=0.945.

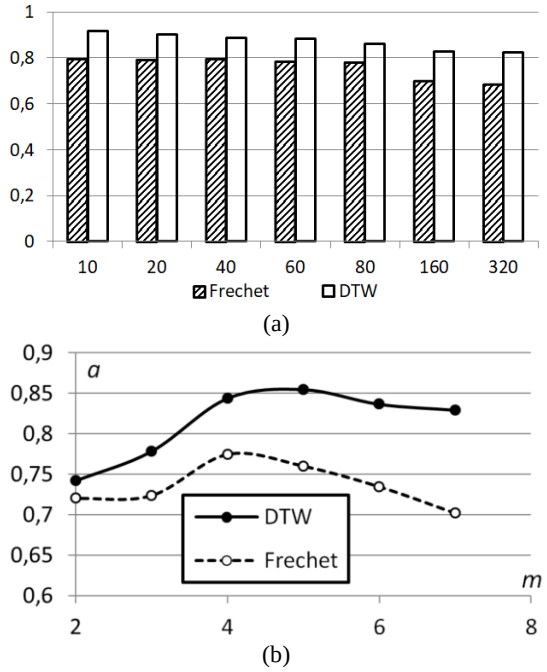

Figure 1. Dependence of the classification accuracy a on the time step s of movement trajectories (a, in frames); on the pose space dimension m for small dimensions (b).

Reducing the time step to s=10 and s=5 allows increasing the accuracy to a=0.967 and a=0.971, respectively, for the dtwi distance. However, this approach leads to a significant increase in processing time due to both an increase in the pose space dimension (48 versus 5) and an increase in the sequence length due to a decrease in the step. For this reason, the paper proposes an alternative approach to the formation of a reduced pose space based on supervised feature selection.

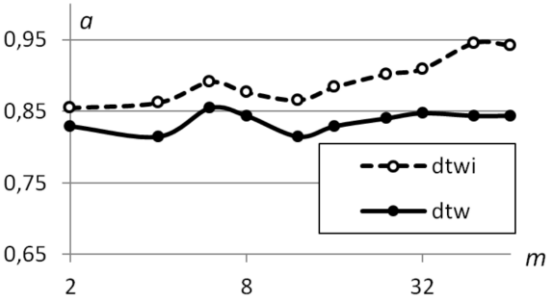

Figure 2. Dependence of classification accuracy a on the pose space dimension m (logarithmic scale) for multidimensional dtw and dtwi distances, step s=20.

4. Supervised Feature Selection

Unlike the PCA technique, supervised feature selection methods use information about the distribution of instances by classes. Such information can be used to evaluate individual features or their combinations. The most obvious strategy is a complete enumeration of all possible combinations of features, but it requires an enormous amount of time. Therefore, in practice, they are satisfied with searching for a suboptimal set of features.

Various approaches can be used to find a suboptimal set of features including sequential addition and elimination of features (Ferri et al., 1994), recursive methods of feature selection, as well as metaheuristic methods of feature selection based on genetic algorithms (Siedlecki and Sklansky, 1989), particle swarm algorithms (Xue et al., 2014), ant colony (Al-Ani, 2005), cuckoo search (Pereira et al., 2013), etc. In all cases, such methods require significant computational costs due to the multiple repetition of training and quality assessment stages, but at the same time ensure the finding of a suboptimal set of features.

In this paper, we will use the forward feature selection (FFS) method based on the sequential addition of features, adapting it for use with an independent version of the DTW distance (dtwi). This option makes it possible to significantly speed up the process of feature selection. Let us consider the FFS method in more detail.

Let F be a set of features defined by their integer indices: $F = \{1, 2,\dots M\}$, where M is the number of features. Let $X^{train} = \left\{x_i^{train}\right\}_{i=1..N_{train}}$ and $X^{val} = \left\{x_i^{val}\right\}_{i=1..N_{val}}$ be training and validation sets of size N_{train} and N_{val} respectively, and $Y^{train} = \left\{y_i^{train}\right\}_{i=1..N_{train}}$ and $Y^{val} = \left\{y_i^{val}\right\}_{i=1..N_{val}}$ be the corresponding class labels. The first two of the indicated sets contains sequences of poses $x_i = x_i^{[1]}, x_i^{[2]},\dots x_i^{[T_i]}$, with the lengths of the sequences T_i being different. As before, we denote by $x_{i,k} = x_{i,k}^{[1]}, x_{i,k}^{[2]},\dots x_{i,k}^{[T_i]}$, k=1..M the partial sequences of the k-th coordinates (features) of the original sequences. Let $D = \{D_1, D_2,\dots D_M\}$ be the pre-calculated matrices of partial one-dimensional DTW distances between the validation and training samples:

$$D_k = \left\{d_{i,j}\right\}_{\substack{i=1..N_{val}\\ j=1..N_{train}}} = \left\{\mathrm{dtw}\left(x_{i,k}^{val}, x_{j,k}^{train}\right)\right\}_{\substack{i=1..N_{val}\\ j=1..N_{train}}}.$$

Let $R = \left\{r_{i,j}\right\}_{\substack{i=1..N_{val}\\ j=1..N_{train}}}$ be a matrix of current distance estimates between the validation and training samples. Initially,

$$r_{i,j} = 0,\ i = 1..N_{val},\ j = 1..N_{train}.$$

Then the feature selection algorithm that forms a set Q containing m features can be represented as the following pseudocode.

$Q = \{\ \}$
while $(|Q| < m)$:
  $k_{best} = -1;\quad a = -1$
  for each $k \in F$:
    $\tilde{R} = R + D_k$
    $\tilde{a} = \mathrm{accuracy}\left(\tilde{R}, Y^{train}, Y^{test}\right)$
    if $(\tilde{a} > a)$:
      $k_{best} = k;\quad a = \tilde{a}$
  $R = R + D_{k_{best}}$
  $Q = Q \cup \{k_{best}\}$
  $F = F \setminus \{k_{best}\}$

In the given algorithm, the output feature set Q is formed by successively adding features k_{best}, which provide the greatest increase in the classification quality a. As the algorithm is executed, the initial feature set F is reduced, and the matrix of current distance estimates R is updated by adding to it the pre-calculated matrix of partial distances $D_{kbest.}$

The classification accuracy score is calculated using the matrix of current distance estimates in the obvious way:

$\text{accuracy}\left(\tilde{R}, Y^{train}, Y^{val}\right)$:

$\text{for } i = 1...n_{val}:$

$\tilde{y}_i^{val} = y_{\tilde{j}}^{train}, \text{ where } \tilde{j} = \arg\min_j \left\{\tilde{r}_{i,j}\right\}$

$\text{return } \frac{1}{n_{val}} \sum_{i=1}^{n_{val}} \mathrm{I}\left(\tilde{y}_i^{val} = Y_i^{val}\right)$

The time-consuming component of the method is the calculation of matrices D_i, i=1..M, containing DTW distances between pairs of validation and training one-dimensional sequences. Such a calculation is performed once for each matrix.

When using the original test set for validation with the described feature selection technique, error-free classification is achieved quite easily (see Fig. 3).

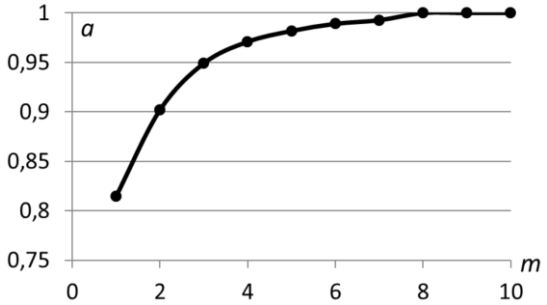

Figure 3. Dependence of the classification accuracy a on the number of selected features m for the feature selection method

For a fairer quality assessment during feature selection, we divided the original training set of the dataset into new training and validation sets. The original test set was used exclusively to assess the quality of the resulting solution.

To divide the original training set, we used the same principles as in the original dataset. As the original training set contained sequences for 7 individuals, the new training set included all possible sequences of P individuals from the original training set, and the validation set included sequences of the remaining (7-P) individuals. We formed the training and validation sets for all possible $\sum_{P=1}^{4} C_7^P$ partitions (here we limited ourselves to considering up to 4 people in the set, thus obtaining 98 partitions) and averaged the accuracy values for these partitions.

The dependence of classification accuracy on the number of selected features is shown in Fig. 4. The feature selection procedure was stopped at 18 selected features (after which the validation accuracy started to decrease), providing an average accuracy of 0.981 across all validation sets. The selected set of features made it possible to achieve an accuracy of a=0.982 on the original test set for the time step s=20.

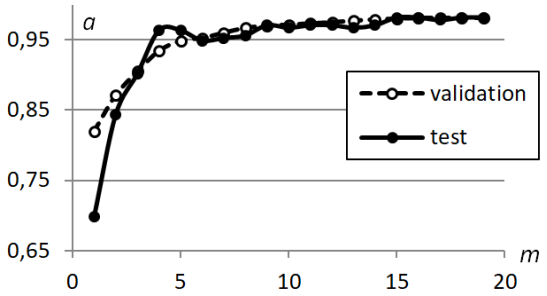

Figure 4. Dependence of the classification accuracy a on the number of selected features m for the validation (averaged values) and test sets.

5. Comparison to Alternative Techniques

In addition to the results obtained using the described approach, we consider the alternative technique based on the subsequence generation approach. We have previously used a similar approach when solving the problem of human action recognition by silhouettes (Shiripova et al., 2020). Here we apply such approach to recognize actions based on motion capture data.

In accordance with this technique, preprocessing similar to the method described above in this paper is first performed. Then, for each sequence of poses describing the motion, a set of subsequences of a given length is generated. Subsequences are generated with some given step, starting from the beginning of the original sequence. Next, for each generated subsequence, a feature vector is formed by concatenating the feature vectors corresponding to individual human poses into a row.

At the classification stage, each of the feature vectors formed for the sequence is assigned a class label indicating the action performed by a person in the original sequence. The set of all pairs consisting of feature vectors and class labels forms a data set for training a given classifier. In this paper, the 1-NN classifier and the support vector machine with a radial kernel (SVM-RBF) were considered as classification methods.

It should be noted that with a sufficiently long subsequence and a large number of sensors, the dimensionality of the generated feature vectors may be high. According to (Strukova et al., 2018), we used the PCA technique to reduce the dimensionality of subsequences space. The transformation parameters were determined based on the training set and were extended to the test set.

When a new test sequence is received, it goes through the feature generation with the subsequent dimensionality reduction stage, after which the decision is made by voting. Here the set of decisions for voting is formed by a previously trained classifier by classifying each subsequence generated for an input test sequence.

After careful selection of parameters (step, subsequence length, subspace dimension, voting scheme), for the 1-NN classifier, the accuracy of a=0.902 was achieved, and for the SVM classifier, the accuracy was a=0.909.

Another considered alternative approach (Ofli et al., 2013), uses the bag-of-features approach. The statistical characteristics (variances) of 21 joint angles calculated for 60 different time windows of original input sequences were used as feature information. On the further details of this approach we refer the reader to the paper (Ofli et al., 2013). It should be noted that the results presented in the paper can be compared, since they were obtained for exactly the same train-test split.

The results of the study of the developed method and the described alternative techniques are summarized in Table 1. As can be seen from the table, the results obtained in the previous sections of this paper significantly outperform the considered alternative techniques.

Method	Accuracy, %
Frechet distance + PCA	79.6
DTW distance + PCA	91.6
DTWI distance + PCA	94.5
DTWI + FFS	98.2
Bag-of-features	
1-NN (Ofli et al., 2013)	74,8
K-SVM (Ofli et al., 2013)	79.9
Subsequence-based classification	
1-NN	90.2
SVM-RBF	90.9

Table 1. Experimental results.

Conclusion

In this paper, we proposed the method for recognizing human actions based on the transition of original motion capture data to sequences of points in a lower-dimensional subspace and subsequent classification of human actions by matching the trajectories of points in the specified subspace. To match the trajectories, we explored the Frechet distance and the Dynamic Time Warping distance in both dependent and independent forms. To form a lower-dimensional pose subspace, we considered known PCA technique and fast supervised feature selection procedure. We compared the proposed technique with two alternatives using open Berkeley Human Action Database and demonstrated its superiority.

Acknowledgements

The study was supported by a grant from the Russian Science Foundation No. 25-21-00413, https://rscf.ru/project/25-21-00413/.

References

Ahad, Md A. R., 2012. Motion history images for action recognition and understanding // Springer Science & Business Media. DOI: 10.1007/978-1-4471-4730-5.

Al-Ani, A., 2005. Feature Subset Selection Using Ant Colony Optimization. *International Journal of Computational Intelligence*, 2(1), 53-58.

Antonucci, A., de Rosa, R., Giusti, A., Cuzzolin, F., 2015. Robust classification of multivariate time series by imprecise hidden Markov models. *International Journal of Approximate Reasoning*, 56(B), 249-263. DOI: 10.1016/j.ijar.2014.07.005.

Aronov, B., Har-Peled, S., Knauer, C., Wang, Y., Wenk, C., 2006. Fréchet Distance for Curves, Revisited. *Algorithms – ESA 2006. Lecture Notes in Computer Science*, 4168, 52–63. DOI: 10.1007/11841036_8.

Bobick, A., Davis, J.W., 2001. The Recognition of Human Movement Using Temporal Templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3), 257-267. DOI: 10.1109/34.910878.

Dhulavvagol, P.M., Kundur, N.C., 2018. Human Action Detection and Recognition Using SIFT and SVM. *Cognitive Computing and Information Processing*. *CCIP 2017. Communications in Computer and Information Science*, 801, 475–491. DOI: 10.1007/978-981-10-9059-2_42.

Du, Y., Wang, W., Wang, L. , 2015. Hierarchical recurrent neural network for skeleton based action recognition. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1110-1118. DOI: 10.1109/CVPR.2015.7298714.

Eiter, T., Mannila, H., 1994. Computing discrete Frechet distance. Technical report.

Ferri, F.J., Pudil, P., Hatef, M., Kittler, J., 1994. Comparative study of techniques for large-scale feature selection. *Machine Intelligence and Pattern Recognition*, 16, 403-413. DOI: 10.1016/B978-0-444-81892-8.50040-7.

Khan, M.A., Javed, K., Khan, S.A. et al., 2024. Human action recognition using fusion of multiview and deep features: an application to video surveillance. *Multimedia Tools and Applications*, 83, 14885–14911. DOI: 10.1007/s11042-020-08806-9.

Kl¨aser, A., Marszałek, M., Schmid, C., 2008. A spatio-temporal descriptor based on 3D-gradients. *BMVC 2008 - 19th British Machine Vision Conference*, 99.1-99.10. DOI: 10.5244/C.22.99.

Kong, Y., Fu, Y., 2022. Human action recognition and prediction: a survey. *International Journal of Computer Vision* 130, 1366-1401. DOI: 10.1007/s11263-022-01594-9.

Mavroudi, E., Bhaskara, D., Sefati, S., Ali, H., Vidal, R, 2018. End-to-end fine-grained action segmentation and recognition using conditional random field models and discriminative sparse coding. *arXiv*:1801.09571. DOI: DOI: 10.48550/arXiv.1801.0957.

Noguchi, A., Yanai, K., 2012. A SURF-Based Spatio-Temporal Feature for Feature-Fusion-Based Action Recognition. *Trends and Topics in Computer Vision. ECCV 2010. Lecture Notes in Computer Science*, 6553, 153–167. DOI: 10.1007/978-3-642-35749-7_12.

O'Hara, A., Lui, Y., Draper, B.. 2011. Unsupervised learning of human expressions, gestures, and actions. *IEEE Int Conf automatic face and gesture recognition*. DOI: 10.1109/FG.2011.5771473.

Ofli, F., Chaudhry, R., Berkeley, Kurillo, G. Vidal, R., Bajcsy, R., 2013. MHAD: A comprehensive Multimodal Human Action Database. *IEEE Workshop on Applications of Computer Vision (WACV)*, 53-60. DOI: 10.1109/WACV.2013.6474999.

Pereira, L.A.M., Rodrigues, D., Almeida, T.N.S., Ramos, C.C.O., Souza, A.N., Yang, X.-S. Papa J.P., 2013. A Binary Cuckoo Search and Its Application for Feature Selection. *Cuckoo Search and Firefly Algorithm. Studies in Computational Intelligence.* 516, 141–154. DOI: 10.1007/978-3-319-02141-6_7.

Pirsiavash, H., Ramanan, D., 2014. Parsing videos of actions with segmental grammars. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 612–619. DOI:10.1109/CVPR.2014.85.

Qazi, H. A., Jahangir, U., Yousuf, B.M., Noor, A., 2017. Human action recognition using SIFT and HOG method. *2017 International Conference on Information and Communication Technologies (ICICT*), 6-10. DOI: 10.1109/ICICT.2017.8320156.

Ryoo, M. S., Aggarwal, J. K., 2006. Recognition of Composite Human Activities through Context-Free Grammar based Representation. *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2006*, 1709-1718. DOI: 10.1109/CVPR.2006.242.

Scovanner, P., Ali, S., Shah, M.A., 2007. 3-dimensional SIFT descriptor and its application to action recognition. *MM '07: Proceedings of the 15th ACM international conference on Multimedia*, 357 - 360. DOI: 10.1145/1291233.129131.

Senin, P., 2008. Dynamic time warping algorithm review. *Information and Computer Science Department, University of Hawaii at Manoa*, 1-23.

Siedlecki, W., Sklansky, J., 1989. A note on genetic algorithms for large-scale feature selection. *Pattern Recognition Letters*, 10(5), 335-347. DOI: 10.1016/0167-8655(89)90037-8.

Simonyan, K., Zisserman, A., 2014. Two-stream convolutional networks for action recognition in videos. *NIPS'14: Proceedings of the 28th International Conference on Neural Information Processing Systems*, 1, 568-576. DOI: 10.5555/2968826.2968890.

Shiripova, L.V., Strukova, O.V., Myasnikov, E.V., 2020. Study of classification techniques for human action recognition based on PCA and width vectors. *IEEE proceedings of 2020 International Conference on Information Technology and Nanotechnology (ITNT)*, 1-4. DOI: 10.1109/itnt49337.2020.9253352.

Srivastava, N., Mansimov, E., Salakhudinov, R., 2015. Unsupervised learning of video representations using lstms. *International conference on machine learning*, 37, 843–852.

Strukova, O.V., Shiripova, L.V., Myasnikov, E.V., 2018. Gait analysis for person recognition using principal component analysis and support vector machines. *IV International Conference on "Information Technology and Nanotechnology" (ITNT-2018). CEUR-WS.org*, 2210. 1-7. DOI: 10.18287/1613-0073-2018-2210-170-176.

Wang, H., Kl¨aser, A., Schmid, C., Liu, C.-L., 2013. Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision*, 103(1), 60–79. DOI: 10.1007/s11263-012-0594-8.

Wang, H., Schmid, C. 2013 Action recognition with improved trajectories. *Proceedings of the IEEE international conference on computer vision*, 3551–3558. DOI: 10.1109/ICCV.2013.441.

Willems, G., Tuytelaars, T., Gool, L., 2008. An efficient dense and scale invariant spatio-temporal interest point detector. *Computer Vision – ECCV 2008. Lecture Notes in Computer Science*, 5303, 650–663. DOI: 10.1007/978-3-540-88688-4_48.

Xue, B., Zhang, M., Browne, W.N., 2014. Particle swarm optimisation for feature selection in classification: Novel initialisation and updating mechanisms. *Applied Soft Computing*, 18, 261-276. DOI: 10.1016/j.asoc.2013.09.018.