Advances in Intelligent Systems Research, volume 174

Proceedings of the 8th Scientific Conference on Information Technologies for Intelligent Decision Making Support (ITIDS 2020)

Stochastic Method for Skeleton Based Human Action Diagnostics

Yury Egorov*
Institute of Mathematics and Computer Sciences
University of Tyumen
Tyumen, Russia
Sirius University of Science and Technology
Sochi, Russia
y.a.egorov@utmn.ru

Irina Zakharova
Institute of Mathematics and Computer Sciences
University of Tyumen
Tyumen, Russia
Sirius University of Science and Technology
Sochi, Russia
i.g.zakharova@utmn.ru

Alexandr Gasanov
Institute of Mathematics and Computer Sciences
University of Tyumen
Tyumen, Russia

Andrew Filitsin
Institute of Mathematics and Computer Sciences
University of Tyumen
Tyumen, Russia

***Abstract*—In recent years, modeling of human actions and activity patterns for recognition or detection of the special situation has attracted a significant research interest. We present our approach for abnormal human action recognition as a sequence of intermediate states. We propose to decompose each action into a sequence of discrete intermediate states and to present state transitions as a stochastic process. Each state is described with the joint locations of a human skeleton. Actions are described with Hidden Markov Model based on the found states and its interconnections. As a result, we combine our stochastic model of human actions with intermediate states described via skeleton joints. Convolutional Neural Network is employed to learn skeleton features for intermediate state recognition. Viterbi algorithm is employed to find model parameters. We implemented proposed methods in a framework for human abnormal action recognition and tested our approach on two samples: MPII Human Pose Dataset and exam footages.**

Keywords—behavior modeling, human action recognition, Hidden Markov Models, intermediate state sequences, skeleton based actions detection

I. INTRODUCTION

The dynamic object recognition problem is a widely studied topic. It has many important applications such as systems for moving object tracking, congestion of city infrastructure, description of dynamic objects, movement modeling. Of particular importance is the study of recognition and diagnosis problems of the object actions in the video [1-3] associated with the creation of interfaces for human-computer interaction, automated gesture recognition systems, intelligent video surveillance systems. In particular such systems ment to detect abnormal situations.

Surveillance system data describing different situations during the exam are the material of this study. In this case, we define abnormal situations as exam process rules violations such as cheating, using prohibited items, violations by the teachers.

Researchers usually develop models of expected situations via probability methods [4, 5] and machine learning algorithms [6, 7] for situation recognition. In this case, those situations that do not fit into the models are considered abnormal. We assume the types of situations violating the exam process rules are known. So, we need to recognize these particular situations in the footages. Thus, recognition is formulated as the classification problem.

In this work, we develop the combined method for abnormal action recognition using the skeleton model and the stochastic approach for actions describing. The main objective of the study is to examine the effectiveness of our method in terms of recognition accuracy of abnormal situations.

II. RELATED WORK

Researchers define two main approaches to solving the abnormal actions diagnostics problem [1-3]. The first approach assumes that abnormal actions belong to one of the classes defined by experts. It reduces the abnormal actions diagnostics to solving the multiclass classification problem [8, 9]. The second approach assumes that abnormal actions are different from the normal actions. It reduces the abnormal action diagnostics to solving the one-class classification problem [6, 7]. It includes the problem of exploration of normal action features distribution [5].

A lot of object action models rely on low-level features such as optical flow. Wang and Snoussi [6] propose a histogram of optical flow orientation as a descriptor of movements. Yuan, Fang, and Wang [10] use selective histogram of optical flow to get the difference between object movements in crowded scenes. The integral optical flow is used in [11] for construction of motion indicators to describe the motions at a regional level. Moreover, other low-level

models are used such as the bag of visual words [4], the models describing patterns of neighboring points trajectories for human action recognition [12], motion autoencoders computed by Motion DeepNet [7], and Gaussian Mixture Variational Autoencoder [5].

A number of models uses high-level features such as interest points features to describe actions. Particularly, Chelli and Pätzold [13] use acceleration and angular velocity data for falling detection. C. Wang, Y. Wang, and Yuille [8] present spatio-temporal model describing the actions with the bag of visual words. In this case, each word presents relationships between motion of the interest points. In [14] Niebles, Chen, and Fei-Fei also extract interest points describing object actions, but they propose to decompose the actions into the simple segments. This idea is used in stochastic and Recurrent Neural Networks modeling. Particularly, in [15, 16], researchers use Multimodal Decomposable Models to encode intermediate states, and Hidden Markov Models to classify actions. Du, W. Wang, and L. Wang [17] developed the action model using Hierarchical Recurrent Neural Network.

There are different types of methods to solve action diagnostics problem. The first type of methods includes the classic machine learning methods such as Support Vector Machine (SVM) [8, 6], Decision Trees, K-Nearest Neighbors Classifiers [13]. The second type of methods is the deep learning methods. Particularly, Simonyan and Zisserman [9] propose Two-stream Convolutional Network. In this network the first stream processes trajectories of moving objects. The second stream processes static images extracting the context in which objects perform the actions. Feng, Yuan, Lu [18] propose PCANet for extraction of the actions features, and Mixture Gaussian Model for actions classification. Researchers also use combined models. Xu, Yan, Ricci, and Sebe [7] extract actions features using Motion DeepNet and apply the ensemble of SVM for abnormal action recognition. We propose combined method for solving action diagnostics problem. Deep learning methods [19, 20] are used to extract interest points describing intermediate states of objects. Hidden Markov Model is used to describe human actions and to solve the diagnostics problem.

III. METHODOLOGY

A. Formulation of the Problem

Object action diagnostics is solved as a classification problem.

Given the following sets

- $X = \{x: x - \text{object state features}\}$.
- $Y = \{y: y - \text{object state}\}$ where each state $y_i \in Y$ corresponds to the features $x_i \in X$.
- $Z = \{z: z - \text{object action}\}$ where each action $z_i \in Z$ corresponds to the sequence of states $y_i = \{y_i^j\}^K$. K is the number of successive states determining the action. K is set a priori.

We need to find two classifiers. The classifier $F_1: X \to Y$ which associates each description of object states $x \in X$ with the correct object state $y \in Y$. And the classifier $F_2: Y^K \to Z$ which associates each sequence $y_i = \{y_i^j\}^K$ with the correct object action $z \in Z$.

B. Intermediate States Model

The object intermediate state is described using a skeleton model. The skeleton model is a vector $x_i = \langle x_i^1, \dots, x_i^j, \dots, x_i^M \rangle$ where x_i^j is coordinates of the object x_i interest point, the number of interest points M is set a priori for $\forall x \in X$. In this representation the model doesn't represent relations between the points, since they a fixed and the same for $\forall x \in X$. We use the graphical representation for the recognition, that is the picture of the object skeleton. The example of the representation is shown in Fig. 1.

Recognition of intermediate state is performed by the classifier $F_1: X \to Y$ which is the Convolutional Neural Network. The purposed network consists of four building units including three convolutional units and a fully connected unit.

Each convolutional unit consists of the following sequence of layers $C - BN - C - BN - P$ where C is the convolutional layer, BN is the batch normalization layer, P is the pooling layer. After every convolutional layer the $LeakyReLU$ activation function is applied. At each convolutional layer, we use nine filters with size 3×3 pixels. As a pooling strategy, we use max polling with a kernel size of two by two pixels.

The fully-connected layer consists of the following sequence of layers $FL_1 - BN - FL_2$ there FL_1 is the hidden layer of twenty neurons, FL_2 is the output layer. We use $ReLU$ as the activation function.

C, BN, FL_1, and FL_2 layers contain learnable parameters.

C. Object Action as a Sequence of States

We present the object action as a stochastic process. In the process the observed object changes his current state to the next one with some probability.

Then we could present the action using Markov Model F_2 defined by tuple $(Y, Z, \mathcal{R}, \mathcal{P}, \mathcal{P}_0, F_1)$ where

- Y is a set of object states.
- Z is a set of object actions.
- $\mathcal{R}$ is a probability distribution of an action $z^{j+1} \in Z$ which is performed by an object in a state y^{j+1}.
- $\mathcal{P}$ is a probability distribution of an object transition from a state y^{j+1} to a state y^j.
- $\mathcal{P}_0$ is a probability distribution of an initial state y^0.

- $F_1: X \to Y$ is a classifier associating each description of object states $x \in X$ with the correct object state $y \in Y$.

Finding a classifier F_2 means finding model parameters $\mathcal{R}$, $\mathcal{P}$, $\mathcal{P}_0$, F_1 with given Y, Z.

It is important to note the constraints for this model.

- It is sufficient to use the same and fixed discretization step of actions for various actions describing.
- We don't consider situations in which observed objects interact with each other.
- At every moment, we have a full object description sufficient for object state recognition. The sufficiency is dermined by an expert.

D. Action Recognition Algorithm

The combined method using the described models is implemented by the following algorithm.

- An intermediate states classifier F_1 is constructed.
- In step $j = 0$, a classifier with the given a classifier F_2, and initial state $y^0 \sim \mathcal{P}_0$.
- In step $j = \overline{1, K-1}$ do the following.
  - The classifier F_1, by description, x^j finds a set $Y^j \subset Y$ from which it chooses a next possible state y^{j+1}.
  - The classifier chooses a next state $y^{j+1} \sim \mathcal{P}(\cdot \mid y^j, Y^j)$.
  - The classifier chooses an action $z^{j+1} \sim \mathcal{R}(\cdot \mid y^{j+1})$ performing by an object.
  - The classifier returns an action class z^{j+1} as a result.

Fig. 1-2 depict the examples of situations, which we need to recognize based on the proposed model. According to expert rules, when detecting cheating, we have to fix that before lowering his hand for the cheat sheet, the student raises his head in order to look around and check if someone is watching him.

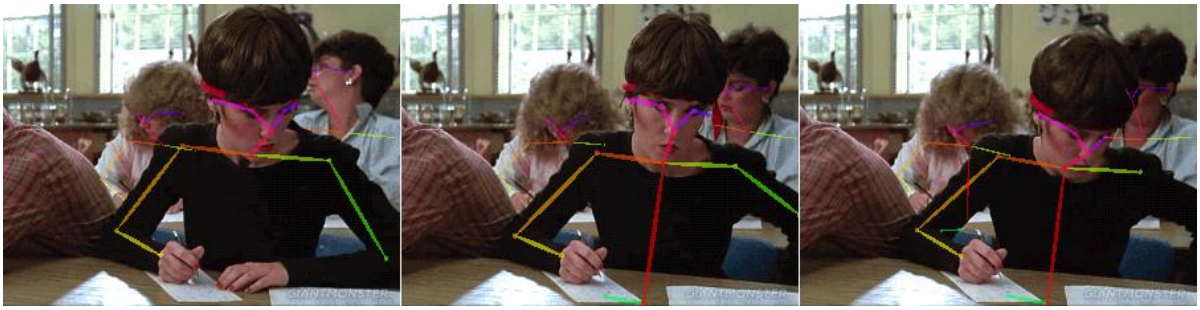

Fig. 1. Storyboard of the first situation. In this case the person is trying to cheat. Frames are ordered from left to right. In the first frame the head is down, and the hands are lying at the desk. In the second frame the head is up, and the left hand is under the desk. In the third frame the head is down, and the left hand is under the desk.

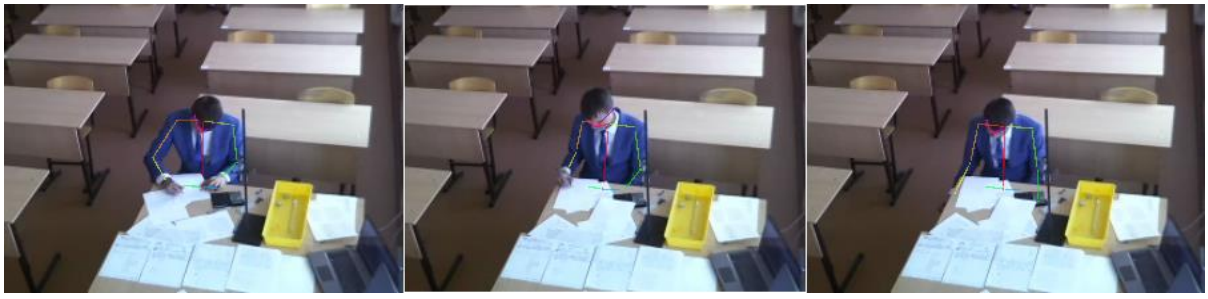

Fig. 2. Storyboard of the second situation. In this case the person doesn't cheat but behaves in a similar way. Frames are ordered from left to right. In the first frame the head is down, and the hands are lying at the desk. In the second frame the head is down, and the left hand is under the desk. In the third frame the head is down, and the both left and right hands are under the desk.

IV. EXPERIMENTS

We conduct two experiments to evaluate practical applicability of the proposed method. In the first one, we compare the accuracy of objects intermediate states classification based on the different data representations. The experiment is conducted on MPII Human Pose Dataset [21]. This dataset consists of images of people performing different actions. In the second one we compute the parameters of Hidden Markov Model presenting the state sequences of the interest point on the footages. For this experiment we use exam footages. From the footages we extract the object interest point movement data during the performing of various actions.

A. Comparison of the Data Representations for Object States Classification

We compare three different data representations for solving the objects intermediate states classification problem.

1) Raw images of the objects.
2) Images of the objects with the skeleton model in graphical representation.
3) The skeleton model in graphical representation only.

Convolutional Neural Network is used as intermediate states classifier F_1. The classification accuracy is used for comparison of the representations.

We use MPII Human Pose Dataset [21] for the experiment. All the images are divided into action categories. Every category is divided into activity types. We choose the five following action categories: fishing and hunting, lawn and garden, water activities, dancing, music playing. We use near 4200 images. We divided the dataset in the following way: 70% are train images, 15% are validation images, 15% are test images. Table 1 shows the results of the experiment. The skeleton model in graphical representation always achieves the lowest accuracy. The images of the objects with the skeleton model get the less accuracy for the fishing and hunting, and music playing categories. At the same time the skeleton model let us improve the accuracy for the water activities category. The accuracy matrices at Fig. 3 show the relationship between the classification accuracy using skeleton model only and the classification accuracy dynamics which we gain by combining the skeleton model with the raw images. Particularly, we can watch that skeleton model doesn't allow us to classify the fishing and hunting, and music playing categories with the necessary accuracy. It takes place because the context of the image is more important for classification than the object position. Therefore, in this case we get decrease of the classification accuracy shifting the classifier attention to the object position with the skeleton model. But for the water activity category, the use of the skeletal model allows us to more accurately distinguish the actions of this category from the actions of the hunting and fishing category. It takes place because a subset of the images in these categories have the same context (i.e. the water activity).

B. Seraching for Action Model Parameters

In this experiment, we find the parameters of the action model F_2. The parameters are presented by the Markov Model transition matrix.

To find out the parameters we use exam footages. From the footages, we obtain the interest point movement data in the following way.

1) The one interest point of the object is selected.
2) The conventional gravity center Q is determined based on the position all interest points of the object.
3) The point trajectories are obtained during the object performing actions.
4) The key positions of the interest point are selected. These positions define five classes: 0 – the point is above the conventional gravity center, 1 – the point is the left of gravity center, 2 – the point is below the gravity center, 3 – the point is the right gravity center, and 4 – the point is in the neighborhood of the gravity point.

The sequences of object states from the interest point positions are the input data for determining the model parameters. Viterbi algorithm [22] is employed to find model parameters. Fig. 4 presents transition matrix of the trained model. The intersection of a row i and a column j contains the transition probability from a state i to a state j.

We assume the selection of the point Q is important but this issue requires additional research.

TABLE I. ACCURACY CLASSIFICATION OF THE INTERMEDIATE STATES

State Type	Raw Images	Raw Images with Skeleton Model	Skeleton Model Only
Fishing and Hunting	**0.62**	0.543	0.161
Lawn and Garden	0.887	**0.909**	0.381
Water Activities	0.659	**0.818**	0.506
Dancing	0.843	**0.885**	0.5
Music Paying	**0.809**	0.607	0.36
Total	0.751	**0.808**	0.414

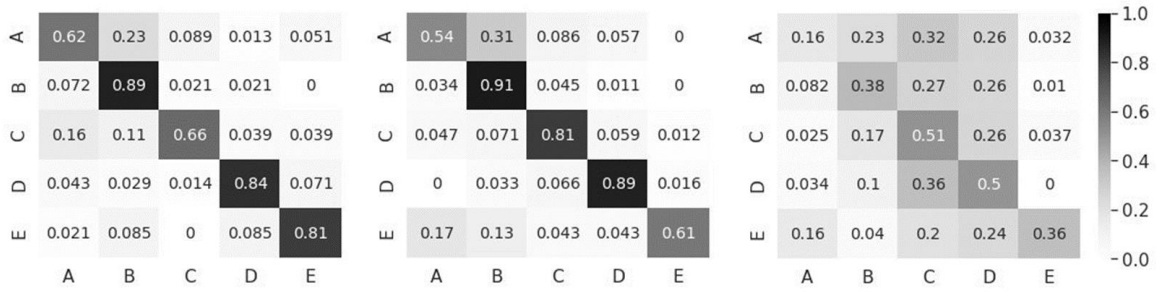

Fig. 3. The confusion matrices for intermediate states classification are ordered left to right. 1 — matrix of classification with the raw images. 2 — matrix of classification with the raw images with the skeleton model in graphical representation. 3 – matrix of classification with the skeleton model in graphical representation only.

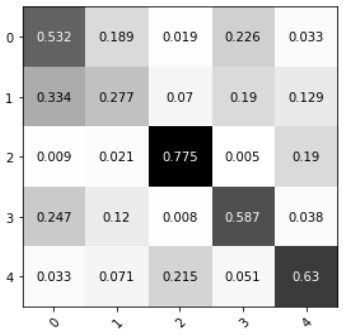

Fig. 4. The transition of the trained model.

V. Conclusion

The exploration of the stochastic method for recognition of abnormal human actions described using skeleton model show certain patterns.

- We get low classification accuracy using the skeleton model only. The result can be explained by the fact that the object context makes significant impact. An object environment and used tools are define the context. We assume that in the case of being objects in the same context, the skeleton model can help to improve the recognition accuracy. In the case of the exam process rules violations recognition problem, the context doesn't have much impact. So, probably, we can use the skeleton model for object states recognition.
- Convolutional Neural Network allows us to use interest points coordinates for solving intermediate states recognition problem without normalization.
- The position of the conventional gravity center of an object depends on the characteristics of recognizable abnormal actions. Therefore, it is necessary to further study the methods for a given point to obtain representative reference point trajectories.

The proposed method relies on a combined model alloying skeletal and stochastic approaches. This fact determines the model flexibility. In particular, it is allowed to use another type of intermediate states classifier or another representation of the source data. At the same time, we can't say that the model is universal, and its use for specific cases requires preliminary research. Of particular interest is the development of the proposed method for identifying abnormal situations associated with the interaction of objects.

Acknowledgment

The reported study was funded by RFBR, project number 19-37-51028.

References

[1] O. P. Popoola and Kejun Wang, "Video-Based Abnormal Human Behavior Recognition—A Review," IEEE Trans. Syst. Man, Cybern. Part C (Applications Rev.), vol. 42, no. 6, pp. 865–878, Nov. 2012.

[2] A. Ben Mabrouk and E. Zagrouba, "Abnormal behavior recognition for intelligent video surveillance systems: A review," Expert Syst. Appl., vol. 91, pp. 480–491, Jan. 2018.

[3] C. Dhiman and D. K. Vishwakarma, "A review of state-of-the-art techniques for abnormal human activity recognition," Eng. Appl. Artif. Intell., vol. 77, pp. 21–45, Jan. 2019.

[4] M. Javan Roshtkhari and M. D. Levine, "An on-line, real-time learning method for detecting anomalies in videos using spatio-temporal compositions," Comput. Vis. Image Underst., vol. 117, no. 10, pp. 1436–1452, Oct. 2013.

[5] Y. Fan, G. Wen, D. Li, S. Qiu, and M. D. Levine, "Video Anomaly Detection and Localization via Gaussian Mixture Fully Convolutional Variational Autoencoder," unpublished.

[6] Tian Wang and H. Snoussi, "Detection of Abnormal Visual Events via Global Optical Flow Orientation Histogram," IEEE Trans. Inf. Forensics Secur., vol. 9, no. 6, pp. 988–998, Jun. 2014.

[7] D. Xu, Y. Yan, E. Ricci, and N. Sebe, "Detecting anomalous events in videos by learning deep representations of appearance and motion," Comput. Vis. Image Underst., vol. 156, pp. 117–127, Mar. 2017.

[8] C. Wang, Y. Wang, and A. L. Yuille, "An Approach to Pose-Based Action Recognition," in 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, pp. 915–922, Jun. 2013.

[9] K. Simonyan and A. Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos," in Advances in Neural Information Processing Systems 27, pp. 568–576, Dec. 2014.

[10] Yuan Yuan, Jianwu Fang, and Qi Wang, "Online Anomaly Detection in Crowd Scenes via Structure Analysis," IEEE Trans. Cybern., vol. 45, no. 3, pp. 548–561, Mar. 2015.

[11] H. Chen, S. Ye, O. Nedzvedz, S. Ablameyko, and Z. Bai, "Motion Maps and Their Applications for Dynamic Object Monitoring," Pattern Recognit. Image Anal., vol. 29, no. 1, pp. 131–143, Jan. 2019.

[12] G. Garzón and F. Martínez, "A Fast Action Recognition Strategy Based on Motion Trajectory Occurrences," Pattern Recognit. Image Anal., vol. 29, no. 3, pp. 447–456, Jul. 2019.

[13] A. Chelli and M. Patzold, "A Machine Learning Approach for Fall Detection and Daily Living Activity Recognition," IEEE Access, vol. 7, pp. 38670–38687, Mar. 2019.

[14] J. C. Niebles, C.-W. Chen, and L. Fei-Fei, "Modeling Temporal Structure of Decomposable Motion Segments for Activity Classification," in 11th European Conference on Computer Vision, Heraklion, Crete, Greece, pp. 392–405, Sep. 2010.

[15] D. Zhang, D. Gatica-Perez, S. Bengio, and I. McCowan, "Semi-Supervised Adapted HMMs for Unusual Event Detection," in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 1, San Diego, CA, USA, USA, pp. 611–618, Jun. 2005.

[16] J. Hendryli and M. I. Fanany, "Classifying abnormal activities in exam using multi-class Markov chain LDA based on MODEC features," in 2016 4th International Conference on Information and Communication Technology (ICoICT), Bandung, Indonesia, pp. 1–6, May 2016.

[17] Yong Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, pp. 1110–1118, Jun. 2015.

[18] Y. Feng, Y. Yuan, and X. Lu, "Learning deep event models for crowd anomaly detection," Neurocomputing, vol. 219, pp. 548–556, Jan. 2017.

[19] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional Pose Machines," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp. 4724–4732, Jun. 2016.

[20] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp. 1302–1310, Jul. 2017.

[21] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D Human Pose Estimation: New Benchmark and State of the Art Analysis," in 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, pp. 3686–3693, Jun. 2014.

[22] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," IEEE Trans. Inf. Theory, vol. 13, no. 2, pp. 260–269, Apr. 1967.