Human Activity Interpretation using Evenly Distributed Points on the Human Hull

Łukasz Kamiński, Krzysztof Kowalak, Paweł Gardziński,
and Sławomir Maćkowiak

Poznań University of Technology, Poland
{lkaminski,kkowalak,pgardzinski,smack}@multimedia.edu.pl,
http://www.multimedia.edu.pl

Abstract. In this paper, a human activity recognition system which automatically detects human behaviors in video is presented. The solution presented in this paper uses a directed graphical model with proposed by the authors Evenly Distributed Points (EDP) method. The experimental results prove efficient representation of the human activity and high score of recognition.

1 Introduction

Behavior understanding and activity recognition using visual surveillance involves the most advanced and complex research in the field of image processing, computer vision and artificial intelligence. The system developed at Poznań University of Technology exploited a directed graphical model based on propagation nets, a subset of dynamic Bayesian networks approaches, to model the behaviors together with Evenly Distributed Points (EDP) method.

There are various representations commonly used for activity understanding, including object-based features, pixel-based features and other feature representations. Object-based representation constructs a set of features for each object in a video based on an assumption that individual objects can be segmented reasonably well for event reasoning. These features include trajectory or blob-level descriptors such as bounding box and shape. Among these features, a trajectory-based feature is commonly employed to represent the motion history of a target in a scene [3,4].

Following the paradigm of object-based representation, there are also attempts to exploit multiple abstract levels of features in a unified framework. For example, Park and Trivedi [5] introduce a framework that switches between trajectory-based features (e.g. velocity and position) and blob-based features (e.g. aspect ratio of bounding box and height) based on the visual quality of detected objects. The system proposed by Poznań University of Technology belongs to the above mentioned category of representation and is a subset of Bayesian networks [2,15]. In the proposed solution, the characteristic features of a human body placed on the human hull and the information about tracking these points are combined together.

Detailed review of other approaches such as Petri nets, syntactic approaches or rule-based approaches can be found in a survey by Turaga et al. [1].

This paper is divided into 4 main sections. Section 2 presents the behavior description and provides detailed explanation on the Evenly Distributed Points on the human hull. Section 3 presents behavior testing system and explains required blocks of video processing. Section 4 presents the assumptions of the experiments and achieved results for exemplary behavior that is a *call for help*.

2 Behavior Description

Let us consider a system based on a single stationary camera that records a scene. Events that happen in that scene can be interpreted (on the frame level) as a sequence of states in time (Fig. 1). Subsequently, those states can be described using the entire picture captured by the camera. However, that amount of information is not necessary, therefore feature extraction techniques are recommended. SIFT (Scale Invariant Feature Transform), Haar-like features and HOG (Histogram of Oriented Gradients) [6] belong to the most popular techniques for feature detection and description. Nevertheless, results shown in the latter paragraph prove that method described in this paper gives equally good results as the above mentioned methods and requires less computational effort. Proposed approach creates key body points on a human hull that can be used to detect human behavior as well.

The behavior of a person can be described by a set of trajectories of characteristic points, as shown in Fig. 2. A set of characteristic points at a given time defines a pose. A set of poses defined for consecutive time points or a set of time vectors for individual points forms a descriptor. Therefore, there are two possible ways to describe a behavior:

- Time-cut interpreted as a sequence of poses in time,
- Space-cut interpreted as a set of trajectories of tracked characteristic points.

In this paper we employ the Time-cut interpretation which correlates to HMM as a representation model.

The set of points to define a pose may have different configuration for different types of behavior to be detected. In other words, for each type of behavior, a

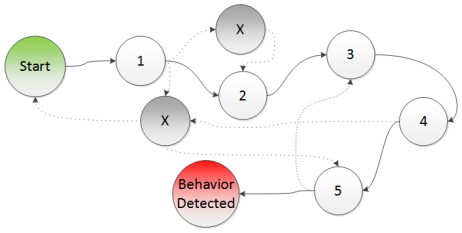

Fig. 1. Behavior as a sequence of poses (states).

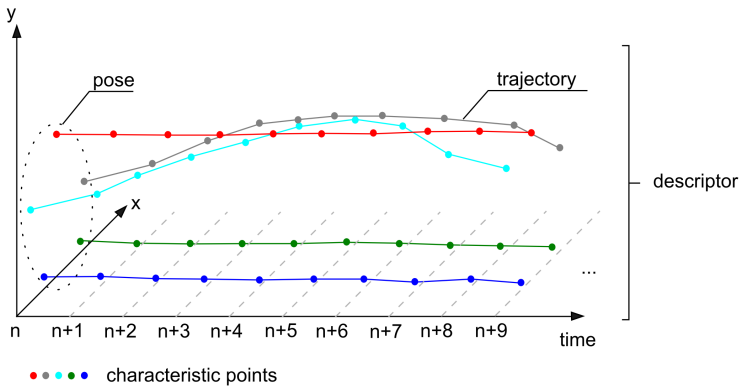

Fig. 2. Example of a behavior as a set of poses, characteristic points form trajectories in time. Trajectory acquired from $Callingforhelp$ behavior.

set of points is generated having a configuration different from the set of points for another type of behavior. For consecutive frames, the positions of points belonging to the Time-cut set are traced and form poses.

The method of selection of points proposed by the authors is called Evenly Distributed Points (EDP). Proposed approach is based on selection of evenly spread, arbitrary number of points from the hull. This kind of approach allows for gradual selection of the level of hull mapping. The method has been depicted in Fig. 3, wherein the following references refer to:

- $Step$- step of selection of consecutive points;
- $Buff$- a buffer storing reminder of a division in order to minimize an error caused by calculations on integer numbers. It is a case when the $Step$ is not an integer;
- $rest(lPktKont/lPkt)$- a function calculating a reminder of a division.

The selected hull points define a pose descriptor of the hull (typically of a silhouette of a person) in a given video frame and are buffered in an output vector as shown in Fig. 3.

More specifically, the procedure in Fig. 3 starts from step 301, where the $Step$ variable is set as value of $lPktKont/lPkt$, wherein $lPkt$ denotes chosen number of equally distributed points on object contour and $lPktKont$ denotes a total number of points in object contour, the $Buff$ variable is set to $rest(lPktKont/lPkt)$ and the i variable is set to 0. Next, at step 302, it is verified whether the value of the $Buff$ variable is less than 1. In case it is not, at step 303 the value of $Buff$ variable is decreased by 1.0 before moving to step 304. Otherwise, the step 303 is skipped and step 304 is executed where the value of the Buff variable is increased by the rest of the quotient $(lPktKont/lPkt)$. Subsequently, at step 305, the i-th point of the hull is added to an output vector as a selected point. Next, at step 306, the variable i is set to a value of $Step + Buff$. Next, it is verified 307 whether i is lower that $lPktKont$ and in

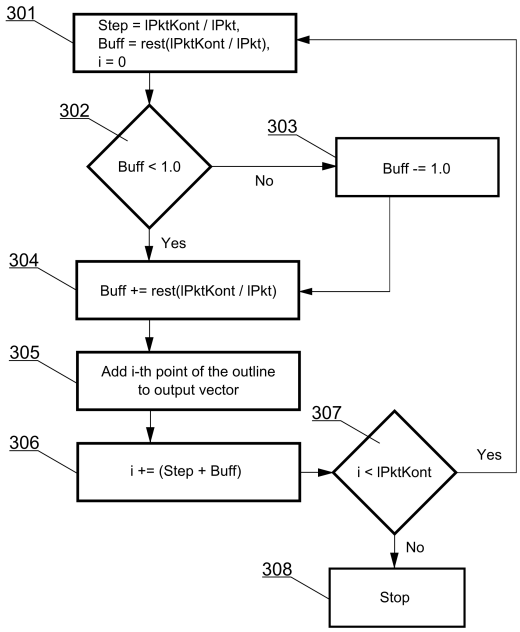

Fig. 3. Evenly Distributed Points - algorithm block diagram.

case it is the process returns to step 301 in order to process the next point or otherwise the procedure ends at step 308.

3 Behavior Detection System

Figure 5 presents the human activity recognition system according to the proposed solution. The method is utilized for analyzing behavior in an intelligent surveillance system. The system is feasible to provide a series of consecutive images (start point 401) of an area under surveillance for consecutive time points.

Presented system uses a single stationary camera. With this assumption, in majority of cases some part of scene contains a static background, which carries redundant information. Therefore, a background subtraction algorithm is used to extract foreground image (e.g. active areas, moving objects) during step 402. From the available algorithms improved adaptive Gaussian Mixture Model (GMM) for background subtraction [13] has been chosen. It is assumed, that the background model is updated during processing of each frame. This results in the best fit of the background model to changing conditions in the scene. To eliminate shadows of the tracked objects, the algorithm described in [12] which is also used during the step 402 is applied. Next step of the algorithm after shadow elimination is characteristic points extraction (step 403) that generates a set of points defining each moving silhouette on an image. This algorithm

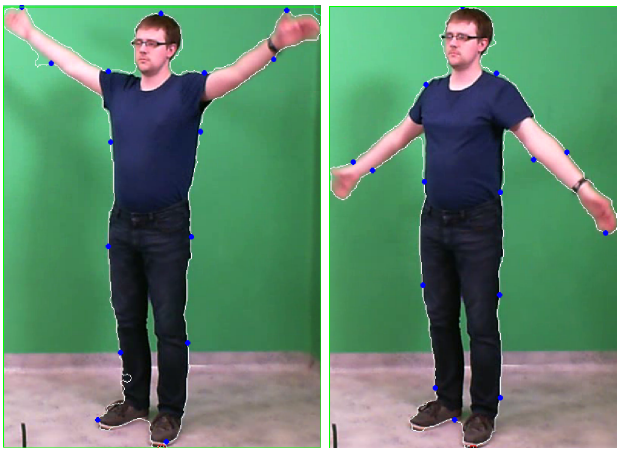

Fig. 4. Characteristics points on the hull.

was described in detail in Section 2. Subsequently, at step 404 a prediction of positions of objects is performed using Kalman Filter [14]. It is necessary because object segmentation is encumbered with some random error. Kalman filtering effectively smoothest trajectories of moving objects. This results in an increase of the accuracy of prediction. Each tracked object has its own Kalman Filter, so that at step 405 the generated trajectories of points for said moving silhouettes can be compared with reference database records of predefined behaviors. The system can use multiple descriptors of the same behavior and enables creation of various combinations of trajectories.

The comparison is performed by calculating the Euclidean distance for pairs of corresponding points. Each pose descriptor (set of characteristic points for a currently processed pose) shall fit within a predetermined range. For example, assuming that 4 points of a person are traced (e.g. two palms and two feet), the characteristic point designated as the "right palm" must be for each frame located in a distance D not larger than δ from the reference right hand for each behavior, namely:

$$D = \sqrt{(x_{obs} - x_{ref})^2 + (y_{obs} - y_{ref})^2} \qquad (1)$$
$$D \leq \delta$$

wherein x_{obs} , y_{obs} designate the position of characteristic points (note: these are not spatial coordinates) and x_{ref} , y_{ref} designate corresponding reference values. Fulfilling the above equation allows a transition to the subsequent state (pose) in a descriptor. Obviously, being in the x state implies the information of visiting a series of previous states (poses) of the analyzed behavior. Thresholds δ were chosen through calibration process with the validation of sequences set. More details about calibration process can be found in Section 4. Thresholds are necessary to deal with different body shapes of humans.

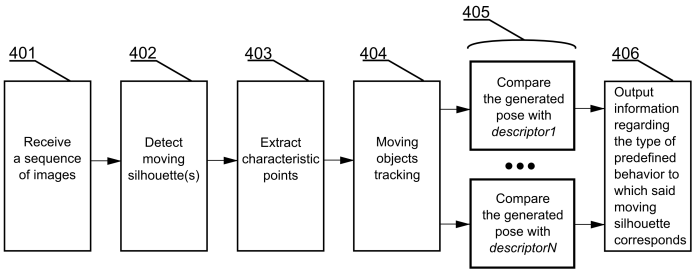

Fig. 5. General diagram of descriptors testing system.

4 Experimental results

The experiments were performed for many different behaviors, such as $Faint$, $Fall$, $Fight$, $StomachPain$ and $Crouch$. In our experiments all sequences in resolution 640x480 were divided with respect to the subjects into a training set (9 persons), a validation set (8 persons) and a test set (9 persons). The classifiers were trained on the training set while the validation set was used to optimize the parameters of each method (δ parameter mentioned in Section 3). The recognition results were obtained on the test set. Authors were looking for well-known databases needed to perform experiments but commonly used test sequences doesn't meet required conditions. This paper contains results concerning $Callforhelp$ behavior.

The efficiency of recognition was analyzed. To evaluate classification algorithms here the precision and recall performance metrics were used that are defined as follows:

$$precision = \frac{TP}{TP + FP} \tag{2}$$

$$recall = \frac{TP}{TP + FN} \tag{3}$$

where TP is the set of true positives, FP is the set of false positives and FN is the set of false negatives.

The effectiveness of the proposed approach was validated using three descriptors that consisted of varied number of states (10, 18 and 36 states) and eight descriptors that represented different realizations of the same human activity with 24 states, together 9 persons from test set.

The experiments were conducted in two steps. In the first step all eleven above mentioned descriptors were tested separately. The results obtained for these descriptors fluctuated between 32% and 79% recall ratio. Subsequently, in the second step the tests concerned varied number of combinations of two or more descriptors. This approach allowed to obtain much better results. Acquired results fluctuated between 58% and 100% recall ratio. These experiments showed that the increase in number of descriptors results in higher recall ratio. The recognition results are presented on Figs. 6, 7 and 8.

5 Conclusions

In this paper, a novel system for human activity recognition from monocular video is proposed. The proposed system is a complex solution which incorporates several techniques which are used to detect moving objects, track them and recognize their activities. This system uses evenly distributed points on human hull which are used in classification process. The results prove that the proposed solution achieves higher detection efficiency with higher accuracy for simultaneous usage of various descriptors while being robust to different weather conditions and behavior realizations. Moreover, the experiment results show that some work regarding characteristic points distribution and their relation to the specific behaviors is worth further research in the proposed approach.

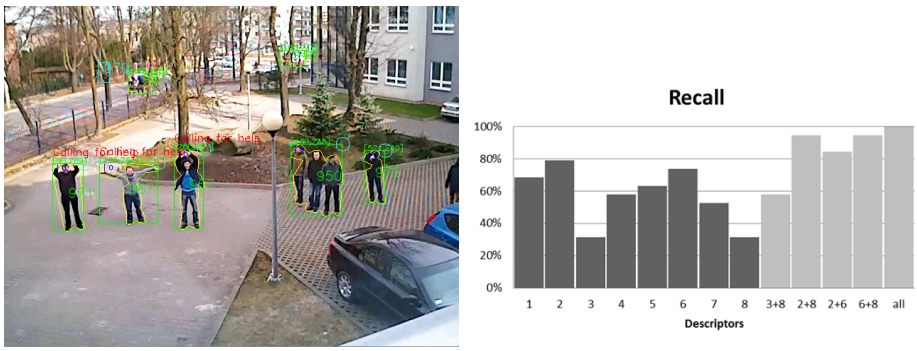

Fig. 6. Left figure: visual evaluation results. Right figure: human activity recognition system evaluation results for eight different descriptors (and their combinations) representing the same human activity.

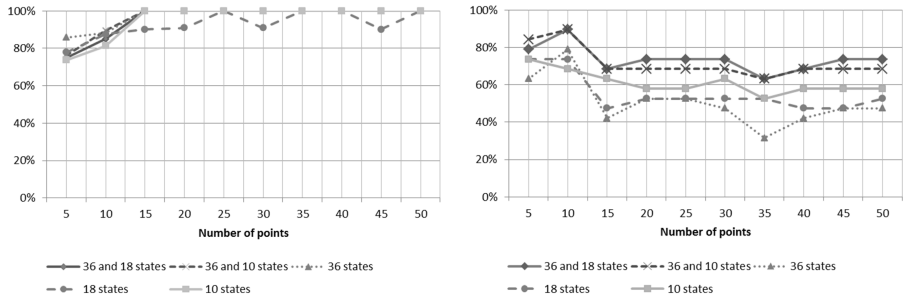

Fig. 7. Human activity recognition evaluation results for descriptors depending on the different number of states, precision (left), recall (right) parameters.

Acknowledgement

Research project was supported by the National Science Centre, Poland according to the Grant no. 4775/B/T02/2011/40.

References

1. P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine recognition of human activities - a survey", IEEE Transactions on Circuits and Systems for Video Technology, 18(11): pp.1473-1488, 2008.
2. Y. Du, F. Chen, W. Xu, and Y. Li., "Recognizing interaction activities using dynamic Bayesian network", International Conference on Pattern Recogntion, pp. 618-621, 2006.
3. Z. Fu, W. Hu, and T. Tan., "Similarity based vehicle trajectory clustering and anomaly detection". In International Conference on Image Processing, pp. 602-605, 2005.
4. N. Johnson and D. C. Hogg. "Learning the distribution of object trajectories for event recognition". Image and Vision Computing, 14(8): pp. 609-615, 1996.
5. S. Park and M. M. Trivedi. "A two-stage multi-view analysis framework for human activity and interactions". In IEEE Workshop on Motion and Video Computing, 2007.
6. N. Dalal; B. Triggs; , Histograms of Oriented Gradients for Human Detection, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Montbonnot, France, 2005.
7. M. Wai Lee; R. Nevatia; "Body Part Detection for Human Pose Estimation and Tracking, IEEE Workshop on Motion and Video Computing (WMVC07), 2007.
8. L. Zhao; "Dressed Human Modeling, Detection, and Parts Localization, The Robotics Institute, Carnegie Mellon University, Pittsburgh, 2001.
9. E. Corvee; F. Bremond; "Body Parts Detection for People Tracking Using Trees of Histogram of Oriented Gradient Descriptors, AVSS 7th IEEE International Conference on Audio Video and Signal Based Surveillance, 2010.
10. K. Mikolajczyk; C. Schmid; A. Zisserman; , "Human Deteciton Based on a Probabilistic Assembly of Robust Part Detectors, T. Pardla and J. Matas (Eds.): ECCV 2004, LNCS 3021, pp. 69-82, 2004.
11. W. Lao; J. Han; P. H.N. de With.; , "Fast Detection and Modeling of Human-Body Parts from Monocular Video, F.J. Perales and R.B. Fisher (Eds.): AMDO 2008, LNCS 5098, pp. 380-389, 2008.
12. R. Cucchiara, C. Grana, G. Neri, M. Piccardi, and A. Prati, The Sakbot System for Moving Object Detection and Tracking, Video-Based Surveillance SystemsComputer Vision and Distributed Processing, pp. 145-157, 2001.
13. Zivkovic Z., "Improved Adaptive Gaussian Mixture Model for Background Subtraction", Proceedings of ICPR, 2004.
14. G. Welch, G. Bishop,: "An Introduction to the Kalman Filter", TR 95-041 Department of Computer Science University of North Carolina at Chapel Hill, 2006.
15. S. Gong and T. Xiang, "Recognition of group activities using dynamic probabilistic networks", IEEE International Conference on Computer Vision, pp. 742749, 2003.