A detection method based on Bayesian hierarchical network for abnormal interaction

Ye Su[1, a], JianXin Song[1, b]

[1]Nanjing University of Post and Telecommunications, Nanjing 210003, China.

[a]suyesean@163.com, [b]songjx@njupt.edu.cn

Keywords: abnormal interactions, feature exaction, Bayesian hierarchical network.

Abstract. Detecting the abnormal human interactions is vital in our daily life, especially when the society pay more attention to public security. But most researches didn't spare enough attention on abnormal interactions. In this paper, salient features are extracted for abnormal interactions, and the amounts of features are reduced to decrease the computation burden. Based on the extracted features, Bayesian hierarchical network is applied to estimating the pose of both persons. Then the corresponding rules for abnormal interaction detection are proposed. Finally, detection results are achieved based on the rules. UT-Interaction dataset is used for experiments. And the results show that the method outperforms with other methods in precision and sensitivity.

1. Introduction

Human action detection has been one of the most important research topics in computer vision. Much progress has been achieved so that many mature techniques have been widely used in various fields such as intelligent video surveillance, human-computer interfaces and video content retrieval. And the accuracy has been considerable high [1].

Due to the former research achieved great success in the single person action, recently many researches has focused on multi-person activities. Rota et al proposed an algorithm on complements motion features with proxemics cues to detect the presence of interactions in surveillance [2]. Taj et al studied Bayesian network-based methods and their variants to analyze interactions in videos [3]. As demonstrated above, we can see that human interaction detection is one of the extraordinarily hot research fields of video content understanding. But most of researches pay much attention to common interactions. In our daily life, we care more about whether the scene is safe. So accurately recognition and detection the abnormal human interaction play a key role in video surveillance in real life. So we proposed the method for abnormal interactions in this paper.

Usually, the interaction detection can be decomposed into three steps: firstly, extracting of the features that can best describe the character interactions; secondly, building and training the interaction models; finally detecting the certain interactions. See Fig. 1 for the flow diagram of abnormal interaction detection.

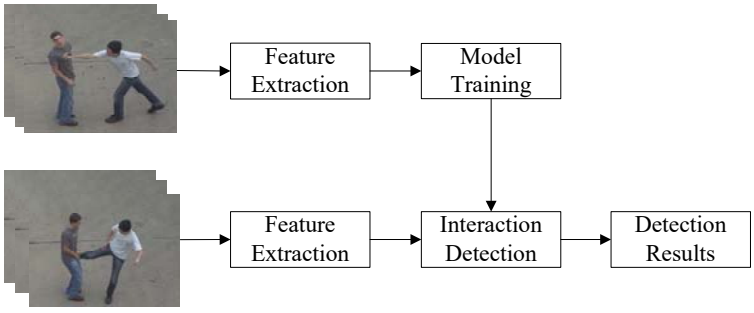

Fig. 1 The flow diagram of common interaction detection

In order to detect and recognize the interactions properly, the performance of extracting object features should be taken into consideration. Many traditional technologies have been used to represent interaction characters for example HOF [4], SIFT [5] and variants. Dalal et al use Histograms of Oriented Gradient (HOG) descriptors to detect and the experiments outperform the

existing technologies [6]. Even though they extracted proper features for detection and recognition, too much features and large amounts data caused too many calculation, especially when we use real time datasets for experiment. As for abnormal interactions, torso features exert important effects on the detection. So in this paper, we proposed a method to extract salient features and reduce the amount of features to reduce the computation burden for abnormal interactions.

Furthermore, in order to understanding the content of the video so that we can detect abnormal interactions properly, modeling the interactions properly makes sense. The existing research works adopt various models. Ivanov and Bobick presented a probabilistic syntactic method to detect and recognize the interactions between multiple agents [7]. They divided the process into two levels. At the lower level, they adopted HMM to recognize atomic actions and the output provided service for the higher level with context-free grammar mechanism. Du et al proposed HDS-DBN (Hierarchical Durational-State Dynamic Bayesian Network) to model interactions between two persons [8]. They decomposed an activity into multiple interactive stochastic processes and each corresponding to one scale of motion details. Park and Aggarwal divide the recognition process into three layers [9]. At low level, body parts are estimated and form into the overall body pose. At middle level, they use DBN to model action of a single person. Based on the relative constraint, they construct a decision tree to recognize two-person interactions. The researches above have shown that decomposing the process of interactions recognition into multiple levels is effective to solve the detection and recognition problems. But they seldom pay attention to abnormal interactions and the really interactions we care in daily life is whether the interactions are abnormal or not. So we adopt the hierarchical methodology to evaluate the abnormal interactions poses in order to reduce the complexity and improve the accuracy. Finally correspondence rules for the interactions are set up to infer and detect the abnormal interactions between two persons.

In this paper, we applied a feature exaction method for abnormal interactions. The Bayesian hierarchical network is optimized to reduce the computation burden. Finally, we proposed rules for detecting abnormal interactions based on the relations between the pose sequence and interactions. Experiment results show that our method obtain good results.

The rest of the paper is organized as follows. Section 2 presents the feature exaction for abnormal. The process of estimate interactions poses using the hierarchical network is also presented. And finally we give out the corresponding rules for abnormal interaction detection. Experimental results will show in Section 3. Section 4 summarizes conclusions and sketches future directions of research.

2. Proposed approach

In this paper, we refer to the action performed by a single person as action, and the action performed by a group people as activity. As a type of activity, human abnormal is defined as an activity that contains two persons in a certain time interval and one touches (or is going to touch) the other resulting in the abrupt reactions of the person being touched. Human abnormal interactions are mainly included with punch, kick and push.

The common process of activity detection is composed of five parts. First, the foreground should be extracted from the image. Since some body parts make sense of the interactions, so certain body parts should be segmented. Then, features of the body parts segmented above can be extracted using ellipses and convex hulls which Park and Aggarwal have conducted in [9]. Afterwards, based on the features we construct a model to estimate pose of the body parts significant to the abnormal interactions. Finally, we set up rules for detecting human abnormal interactions using the estimation results. We can see the process of abnormal interaction detection in the flow diagram in Fig. 2. Now we will discuss the pose estimation for abnormal interactions and the rules we set up for abnormal interaction detection.

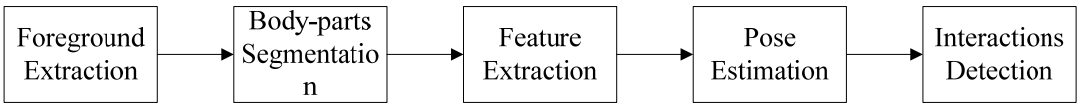

Fig. 2 The process of abnormal interactions detection

2.1 Feature extraction for abnormal interactions.

Since the arms and legs play an important role in abnormal interaction and what is more important is that too many features will cause intractable solution and complex computation. Taking the factors demonstrated above into consideration, we present a method to extract salient features of arms in the upper body and legs in the lower body.

Due to the sensitivity to the outermost points of the convex hull, we use convex polygons to extract the salient features to represent upper body and lower body. We define convex polygon as a curve composed several points in the outermost region, which can encircle the whole region of certain part. The different body part is segmented using the method presented in [9]. We use the algorithm in [10] to determine polygons. We select maximum curvature points as the position of the arm and the leg, the salient features respectively in the upper body and lower body, vital in abnormal interactions.

The directions of upper body and lower body are also important, because different target that the arm stretch to represent different interactions. For example, if the arm tends to touch the head of the other person standing still or the upper body of the other person who tends to move backwards, it is very likely to infer that the interaction is striking (or punch). As a result, we will determine the interaction as abnormal interaction. But if the arm just stretches out and the direction of the upper body is middle, the interaction is likely to be handshake which is normal. So we use an ellipse for each person to describe the direction of the action. The ellipticity and the rotation angle of the ellipse play important role in the features of the interactions. We use principal component analysis (PCA) to obtain the lengths and directions of the long axis and the short axis. The eigenvector can describe the information above. We can see the results of the feature extraction in Fig 3.

Fig. 3 The results of feature extraction. The left is the ellipse representation and the right convex representation.

2.2 Pose estimation for abnormal interactions.

When we extract features, we can construct a model to further extract descriptor of the objects. We use the Bayesian hierarchical network to estimate the pose of upper body and lower body. Usually we just need to perform estimation once a time for one frame, because the person produces abnormal dangerous action with only one part of the whole body at one time. For example, the man usually cannot hit the other one with an arm and a leg at the same time. So we can determine the body part which produces the action to reduce the computational work. The Bayesian network we construct for the upper body pose is shown in Fig 4. Here H_1 is the vertical pose of the upper body. H_2 is the horizontal pose of the upper body. V_1 is index for the vertical position of the candidate salient point of the convex hull. V_2 is index for the ellipticity of the ellipse of the upper body. V_3 is index for the angle rotation of the ellipse. V_4 is the index of horizontal position of the candidate salient point of the convex.

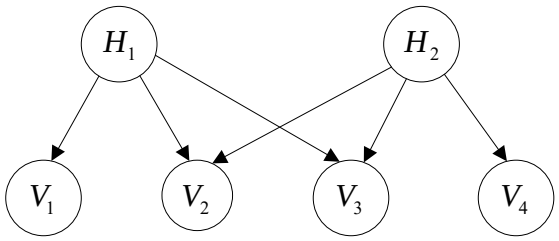

Fig. 4 The Bayesian network for upper body.

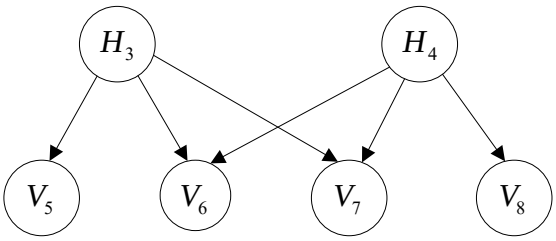

Fig. 5 The Bayesian network for lower body.

In Fig 3, we can see that point A is the salient point of the left person and point B, the right person. Based on the figure 4, we defined the state of hidden node H_1 as

H_1={head，upper in upper body, lower in upper body,putdown}

and the state of hidden node H_2 as

H_2={withdraw，middle, stretch out}.

Moreover, we define the visible node's state as follows.

$V_1 = \{0, 0.2, 0.35, 0.5\}$, $V_2 = \{0.3, 0.6, 0.9\}$

$V_3 = \{30^\circ, 60^\circ, 90^\circ\}$, $V_4 = \{0, 0.1, 0.15, 0.23\}$

The joint probability of the Bayesian network is factored into conditional probabilities and prior probabilities.

$$\begin{aligned} P(V_{1:4}, H_{1:2}) &= P(V_{1:4} | H_{1:2}) P(H_{1:2}) \\ &= P(V_{1:4} | H_{1:2}) P(H_1) P(H_2) \\ &= P(V_1 | H_{1:2}) P(V_2 | H_{1:2}) P(V_3 | H_{1:2}) P(V_4 | H_{1:2}) P(H_1) P(H_2) \\ &= P(V_1 | H_1) P(V_2 | H_{1:2}) P(V_3 | H_{1:2}) P(V_4 | H_2) P(H_1) P(H_2) \end{aligned} \tag{1}$$

Our goal is to estimate the belief of the states of the hidden nodes $H_{1:2}$ given the evidence $V_{1:4}$.

$$\begin{aligned} P(H_{1:2} | V_{1:4}) &= \frac{P(V_{1:4}, H_{1:2})}{P(V_{1:4})} \\ &= \frac{P(V_{1:4}, H_{1:2})}{\sum_{allH_{1,m}} \sum_{allH_{2,n}} P(V_{1:4}, H_{1:2})} \\ &= \frac{P(V_1 | H_1) P(V_2 | H_{1:2}) P(V_3 | H_{1:2}) P(V_4 | H_2) P(H_1) P(H_2)_1}{\sum_{allH_{1,m}} \sum_{allH_{2,n}} [P(V_1 | H_1) P(V_2 | H_{1:2}) P(V_3 | H_{1:2}) P(V_4 | H_2) P(H_1) P(H_2)]} \end{aligned} \tag{2}$$

The conditional probability and the prior probability are obtained from training the video data.

The lower body is similar to the upper body. The Bayesian network is shown in Fig. 5.

The meanings of all the nodes are demonstrated as follows. H_3 is the vertical pose of the lower body. H_4 is the horizontal pose of the lower body. V_5 is index for the vertical position of the salient point of the convex hull. V_6 is index for the ellipticity of the ellipse of the lower body. V_7 is index for the angle rotation of the ellipse. V_8 is the index of horizontal position of the salient point of the convex.

And the states of every node are also shown in the following part.

H_3={putdown,middle of the lower body,high of the lower body}

H_4={withdraw,middle,stretch out}

$V_5 = \{0.5, 0.75, 1\}$, $V_6 = \{0.3, 0.6, 0.9\}$

$V_7 = \{30^\circ, 60^\circ, 90^\circ\}$, $V_8 = \{0, 0.15, 0.3\}$

The joint probability of the Bayesian network is factored into conditional probabilities and prior probabilities.

$$P(V_{5:8}, H_{3:4}) = P(V_5|H_4)P(V_6|H_{3:4})P(V_7|H_{3:4})P(V_8|H_4)P(H_3)P(H_4) \quad (3)$$

Our goal is to estimate the belief of the states of the hidden nodes $H_{3:4}$ given the evidence $V_{5:8}$.

$$P(H_{3:4}|V_{5:8}) = \frac{P(V_5|H_3)P(V_6|H_{3:4})P(V_7|H_{3:4})P(V_8|H_4)P(H_3)P(H_4)}{\sum_{allH_{3,m}}\sum_{allH_{4,n}}[P(V_5|H_3)P(V_6|H_{3:4})P(V_7|H_{3:4})P(V_8|H_4)P(H_3)P(H_4)]} \quad (4)$$

2.2 Detection rules for abnormal interactions.

After extracting the body part descriptors, we can estimate the overall descriptor for the separate person to produce coarse description about the interactions. Now, we can obtain the pose of two persons of the certain frame using the hidden state indices of $H_{1:4}$. To demonstrate clearly, we give an example as follows. The pose of the left person is represented as p_t^1=[$H_{1:4}$]. The right person is similar with the left and we define the pose of the right is p_t^2=[$H_{1:4}$]. We use A, B etc. to represent the states sequentially.

For simplification, we focus on the left person when analyzing. Since the head suffers from hurts easier, so we judge the interaction abnormal for safety's sake. Sometimes, the interactions rely on both of the persons, so we have to take the reaction of the other person into consideration. For example, if the left person raises his arm at the upper part in the upper body and the right doesn't react to the action, we will judge the interaction doesn't cause abnormal effects. So we determine the interaction normal. For details, we can infer coarse description for abnormal interactions from the pose representation based on the rules in the Table 1.

Table 1 Correspondence between the pose sequence and the interactions

Body parts	Sequence left p_t^1	Sequence right p_t^2	Interactions
Upper body $H_{1:2}$	AA, AB, AC	Any	Punch
	BA, BB, BC	Except CA	Push
	CA	AA, AB, AC	Punch
		Others	Normal
	CB, CC	AA, AB, AC	Punch
		BA, BB, BC	Push
		Others	Normal
Lower body $H_{3:4}$	CB, CC	Any	Kick
	Others	CB, CC	
		Others	Normal

3. Experiments

3.1 Dataset.

In order to validate our proposed method effective, we use UT-Interaction dataset for experiments. The datasets contain five classes of interactions (except for point): shake-hands, hug, kick, punch, push. The latter three types of interactions are abnormal interactions. Since the action point is performed by one person, we will not take it into consideration. The dataset is divided into two sets. Both of the two sets are composed of 10 videos. In the segmented sets, there are 10 videos for each interaction. The resolution of the video is 720*480, 30fps [11].

3.2 Experiments and results.

In our experiments we will apply the method to the segmented datasets. In order to obtain the conditional probability and the prior probability used when estimating the belief of the state of the hidden nodes, we choose 70% videos of the two sets for training and the remaining videos are used

for generating results. The training procedure is performed using the method of histogram procedure proposed in [9].

Based on the pose sequence and the interactions in Table 1, we can obtain the experiments results shown in the confusion matrixes in Fig. 6. Vertical ordinate represents the real interactions in the video, and the horizontal ordinate represents the detection results.

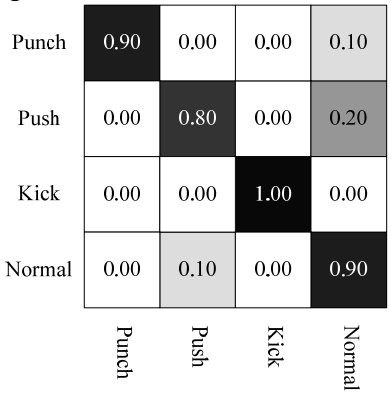

Fig. 6 The confusion matrix of the results of our proposed method

In order to evaluate our proposed method, we use two metrics namely precision and sensitivity. We define the precision as the result of the number of true positive detections divided by the sum of false positive and true positive detections. And the sensitivity is the result of number of true positive detections divided by the sum of true positive and false negative detections. See below for detail.

$$precision = \frac{TruePositive}{TruePositive + FalsePositive} \tag{5}$$

$$sensitivity = \frac{TruePositive}{TruePositive + FalseNegative} \tag{6}$$

Based on the results in Fig 6, we can obtain the precision and sensitivity in Table 2.

Table 2 Detection results of proposed method

Interactions		Precision	Sensitivity
Abnormal	Punch	100%	100%
	Push	88%	80%
	Kick	100%	100%
Total		96.4%	93.1%

Table 3 Compared with other detection work involved with abnormal interactions

Methods	Precision	Sensitivity
Our method	96.4%	93.1%
Xiaofei Ji et al[12]	95.6%	84.6%
Y Kong et al[13]	89.0%	95.3%
Mukherjee et al[14]	96.1%	86%

There are 30 abnormal interactions and 20 normal interactions. We use our proposed method to detect 28 abnormal interactions and 22 normal interactions. So the precision of our method is 96.4% and the sensitivity is 93.1%.

Compared with other algorithm shown in Table 3, our method is superior in the precision and sensitivity. The result is acceptable, because we prefer to determine the normal interactions as abnormal rather than leave out the abnormal interactions. What's more, we set up the rule mainly for the abnormal interactions, so the performance is better when detecting abnormal interactions.

4. Conclusions

In this paper, we applied a hierarchical network feature exaction method for abnormal interactions. Based on the correspondence relations between the estimation poses and the description, we can detect the abnormal interactions. The experiment results have demonstrated our method outperforms the existing methods as for abnormal interactions. However, our method has just applied on the UT interaction dataset. The method is not robust when the environment becomes complex. So in the future, to achieve more robust results, we can exploit more information in the videos.

References

[1] Aggarwal J K, Ryoo M S. Human activity analysis: A review[J]. Acm Computing Surveys, 2011, 43(3):194-218.

[2] Rota P, Conci N, Sebe N. Real Time Detection of Social Interactions in Surveillance Video[M]// Computer Vision – ECCV 2012. Workshops and Demonstrations. Springer Berlin Heidelberg, 2012:111-120.

[3] Taj M, Cavallaro A. Recognizing Interactions in Video[M]// Intelligent Multimedia Analysis for Security Applications. Springer Berlin Heidelberg, 2010:29-57.

[4] Laptev I, Marszalek M, Schmid C, et al. Learning realistic human actions from movies[C]// Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2015:1-8.

[5] Lowe D G. Object recognition from local scale-invariant features[C]// The Proceedings of the Seventh IEEE International Conference on Computer Vision. IEEE, 1999:1150.

[6] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]// IEEE Conference on Computer Vision & Pattern Recognition. 2013:886-893.

[7] Ivanov Y A, Bobick A F. Recognition of Visual Activities and Interactions by Stochastic Parsing[J]. Pattern Analysis & Machine Intelligence IEEE Transactions on, 2000, 22(8):852-872.

[8] Du Y, Chen F, Xu W, et al. Activity recognition through multi-scale motion detail analysis[J]. Neurocomputing, 2008, 71(16–18):3561-3574.

[9] Park S, Aggarwal J K. Segmentation and Tracking of Interacting Human Body Parts under Occlusion and Shadowing[C]// The Workshop on Motion & Video Computing. 2002:105.

[10] Graham R L. An efficient algorith for determining the convex hull of a finite planar set[J]. Information Processing Letters, 1972, 1(1):132–133.

[11] M. S. Ryoo and J. K. Aggarwal. UT-interaction dataset, ICPR contest on semantic description of human activities (SDHA). http://cvrc.ece.utexas.edu/SDHA2010/Human Interaction.html, 2010.

[12] Xiaofei Ji, Changhui Wang. Multiple Feature Voting based Human Interaction Recognition[J]. International Journal of Signal Processing, Image Processing and Pattern Recognition, 2016,1(9):323-334.

[13] Kong Y, Liang W, Dong Z, et al. Recognising human interaction from videos by a discriminative model[J]. Iet Computer Vision, 2014, 8(4):277-286.

[14] Mukherjee S, Biswas S K, Mukherjee D P. Recognizing interaction between human performers using 'key pose doublet'[C]// Proceedings of the 19th ACM international conference on Multimedia. ACM, 2011:1329-1332.

[15] Lan T, Wang Y, Yang W, et al. Discriminative latent models for recognizing contextual group activities.[J]. IEEE Transactions on Software Engineering, 2012, 34(8):1549-62.

[16] Desai C, Ramanan D, Fowlkes C. Discriminative models for static human-object interactions[C]// IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. 2012:9-16.