Violence Recognition Using Harmonic Mean of Distances and Relational Velocity with K-Nearest Neighbour Classifier

Muhammad Alhammami(✉), Chee Pun Ooi, and Wooi-Haw Tan

Faculty of Engineering, Multimedia University, Jalan Multimedia, 63100 Cyberjaya, Malaysia
dr.mhammami@outlook.com, {cpooi,twhaw}@mmu.edu.my

Abstract. Violence recognition falls in the domain of action recognition which has gained considerable attention and importance due to its wide application. Violence recognition has to take place in real time. One main approach to accelerate the recognition is to efficiently choose and calculate suitable features to be used in recognition which is known as feature selection. This paper proposes the use of only nine harmonic means of relational distances between pairs of six joints and one relational velocity between 2 joints. The selected joints are chosen carefully based on having the highest information gain for the recognition. The results show that very high accuracy rate of 99.8 % can be achieved with k-nearest neighbours (k-NN) classifier. This excellent recognition rate would encourage researchers in trying to implement the proposed approach in hardware, as it uses comparatively few data for processing with simple algorithms.

Keywords: Violence recognition, human action recognition, feature extraction · Skeleton, harmonic means, velocity, classification

1 Introduction

Human action recognition, including violence recognition, needs complex operations to extract useful information from image sequences or video. These operations include image acquisition, image pre-processing, features extraction and classification etc. The bigger the image data, the underlying operations will be more complex and the performance will be relatively slow.

Most researchers in the field of action recognition focus on increasing the accuracy rate regardless of the complexity of the data and algorithms. This might be feasible for theoretical researches. These approaches are not suitable for implementation in the embedded systems as they are not optimized for the systems to work efficiently. One viable solution in accelerating the recognition process is to improve the quality of the features calculated in each frame. The usage of skeleton data obtained from depth sensors is very promising in computing features for action recognition. However, using all the skeleton data as inputs will significantly increase the complexity of the system and dramatically decrease the performance.

H. Badioze Zaman et al. (Eds.): IVIC 2015, LNCS 9429, pp. 132–139, 2015.
DOI: 10.1007/978-3-319-25939-0_12

In this paper we present a model of violence recognition system, which only uses nine harmonic means of relational distances between a few pairs of joints together with one relational velocity between two joints as features of classification with k-nearest neighbors (k-NN).

2 Related Works

A lot of efforts have been spent in the field of action recognition. Researchers have to answer which, why and how certain actions should be considered. Generally, actions can be categorized as normal or abnormal, single action or interacted actions. They may take place indoors or outdoors. The purpose of the action recognition system is very important in defining which actions to be considered. These purposes may include healthcare, security or entertainment. There are two main methods for action recognition. The first method is a model-based approach where actions are described on a high level. The second method is learning patterns from training dataset of actions and recognizing new actions using the learned patterns. The steps of building the dataset and recognition are: segmentations, feature extractions and representations and classification [1].

Using skeleton data for action recognition is recently becoming more popular and many dataset have been built using these data. Li *et al*. [2] recognized human actions from series of depth maps. They modeled the dynamics of the actions by using an action graph and characterized a set of salient postures that correspond to the nodes in the action graph using a bag of 3D points. Ni *et al*. [3] presented a publicly releasable human activity video database which contains synchronized colour-depth video streams. The previous two papers contained only depth maps and colour as inputs. Joint sequences from depth sensors were used as a feature by Masood et al. [4] and Sung *et al*. [5]. Only skeleton joints were used by Masood *et al*. [4] as a feature for single activity recognition and actions were detected by logistic regression in real time. Colour, depth and skeleton joints were used by Sung *et al*. [5] as features for daily activities. The hierarchical maximum entropy Markov model (MEMM) was used for classification. Yun *et al*. [6] used synchronized video, depth and motion capture data for creating a human activity dataset about two person interactions. Multiple Instance Learning in a boosting framework (MILBoost) was applied for classification. Liu *et al*. [7] presented a method of recognizing human actions by using Microsoft Kinect sensor, K-means clustering and Hidden Markov Models (HMMs).

From the literature, it can be seen that the use of harmonic mean in action recognition literature has not been investigated yet.

3 Methodology

The methodology consists of six steps. Firstly, we choose a dataset which at least has some violent actions as a benchmark. Secondly, we analyze all actions to determine the most important joints engaged in actions. The third step is the segmentation of all videos. In the fourth step we use the concept of harmonic mean to deal with

irrelative frames to actions. The fifth step is to choose the most relevant features using the information gain measure. The final step is finding the best classification algorithm in terms of accuracy and performance.

The first step in our methodology is to select a benchmark dataset that contain skeleton data for the action that to be used for this project. This dataset has to include the interactions between two persons in performing some aggressive actions. The dataset selected in this project is the dataset "Two-person Interaction Detection Using Body-Pose Features and Multiple Instance Learning" [6]. The action classification of this dataset can be used to verify the method proposed in this project. The chosen dataset as shown in Fig. 1 shows 8 actions and its information about RGB, depth and skeleton of each frame. The actions as mentioned in [6] are: approaching, departing, kicking, punching, pushing, hugging, shaking hands and exchanging. At this point we imported the skeleton data of the dataset and visualize them for further analysis.

Fig. 1. The subject dataset contains RGB, Depth and skeleton data [6].

The outputs of the proposed methodology are geometric relational body-pose features. The main challenge to get bis that there are 15 joints for each person in each frame, the relative Euclidian distances between each 2 different joints have to be calculated. Hence in total there will be 435 values in each frame, and when studying the development of an action during a window of W frames, we will have a matrix of 435xW dimensions. This amount of data is considered huge and difficult to process on personal computers. Therefore, the second step in our methodology is to evaluate each action in order to determine the most important joints involved in each action.

In the third step we manually segment all the scenes we have about the eight actions. And we assign each sliding window to one action depending on the main action in each sliding window. Here, we have defined and included two more actions (Approaching and Departing) to the original 8 actions to better and more accurate assigning each window to one action.

The fourth step in our methodology is to calculate the harmonic mean of all relational Euclidian distances between all pairs of the previously selected joints and their relational

velocities in each sliding window of W frames. Since the harmonic mean of a list of numbers tends to bias strongly toward the least elements of the list, it has the advantage of mitigating the impact of outliers with large values. So we use the harmonic mean of the distances to deal with irrelative frames around the main action in each sliding window where the relational distances between the joints are big. At the same time it aggravate the impact of these distances in the main action time as the joints of the two persons are proximate to each other. We also calculate the relational velocities of all selected joints' movements in each sliding window. The relational velocities in each sliding window is obtained via Eq. (1):

$$\mathcal{V}_{\mathcal{I}_{x_i}\mathcal{I}_{y_i}} = \frac{\max\limits_{incurrentwindow}\left(dist\left(\mathcal{I}_{x_i},\mathcal{I}_{y_j}\right)\right) - \min\limits_{incurrentwindow}\left(dist\left(\mathcal{I}_{x_i},\mathcal{I}_{y_j}\right)\right)}{w_{max} - \mathrm{w}_{min}} \quad (1)$$
$$i,j \in \{\text{Joint1..Joint15}\}$$
$$x,y \in \{\text{person1},\text{person2}\}$$

The fifth step in the methodology is the feature selection. Feature selection depends on considering the most relevant attributes by using the information gain measure. This measure is obtained via Eq. (2):

$$Gain\,(S,A) = Entropy\,(S) - \sum_{v \in Values(A)} \frac{|S_v|}{S} Entropy\left(S_v\right) \quad (2)$$

Where Values(A) is the set of all possible values of the feature A, and $\mathrm{S}_v = \mathrm{s} \in \mathrm{S}\,|\mathrm{A}\,(\mathrm{s}) = v$ for a collection of examples S. The entropy is defined via Eq. (3):

$$\text{Entropy}\,(\mathrm{S}) = \sum_{i=1}^{c} - \mathrm{p}i\log 2\mathrm{p}i \quad (3)$$

Where pi is the probability of S to belong to action class i.

After the features selection is done, the final step in our methodology is to find the best classification algorithm in both accuracy and speed. We evaluated most of the related classification algorithms used before in action recognition literature. The test of each classifier is done by 5 folds cross validation. This means four folds for training and one fold for evaluating. This condition is the same in the reference of the dataset so we can compare both results at the end [6].

4 Experimental Results

Based on our methodology discussed in the previous section, we selected the following joints of each actor as a starting point: Heads, Hands and Feet. We then calculated the relative Euclidian distances between the previous joints in all the frames in a sliding window of W frames in each scene. The harmonic means of all these relative distances were then calculated together with the velocities of these joints during each window. By

doing this we got very large amount of features so we had to select the most informative features for the recognition based on the information gain measure.

As a result of our methodology we get that the harmonic mean of the relative distance between the heads of the actors is the most important feature because it has the highest information gain (0.9 bits), then the velocity of this relational movement of the heads (0.78 bits). This velocity feature is the only velocity feature which has a high information gain. Table 1 shows the top ten features which have highest information gain.

Table 1. Proposed features for action recognition

Feature	Information gain (Bits)
Harmonic mean of the distance between heads of persons A and B	0.90
Relative velocity of the heads of persons A and B	0.78
Harmonic mean of the distance between right foot and right hand of person A	0.55
Harmonic mean of the distance between right hand of person A and right hand of person B	0.53
Harmonic mean of the distance between right foot of person A and right foot of person B	0.51
Harmonic mean of the distance between right foot of person A and head of person B	0.50
Harmonic mean of the distance between right hand of person A and head of person B	0.50
Harmonic mean of the distance between right foot of person B and right hand of person B	0.48
Harmonic mean of the distance between right hand of person B and head of person A	0.47
Harmonic mean of the distance between right foot of person B and head of person A	0.43

Therefore we select these features for recognizing the original eight subject actions plus the two added actions (Approaching and Departing) in the second and third case in the training and classification phases as we will see later in this section. But for the first case in the training and classification we will use all the computed distances (not harmonic means of them) and their velocities.

After that we evaluated many classification algorithms under three cases as explained in Table 2:

Table 2. Results of classification.

Case no. and window size (Frames)	Features	Classifiers	Time model (s)	Average accuracy (%)
Case 1: W = 3	All distances and velocities	Multilayer Perceptron	NA	NA
		NaiveBayes	0.84	58.8
		Random Forest	3.02	79.1
		K-nearest neighbours	0.02	84.6
		Support Vector Machine	65.44	84.8
Case 2: W = Whole Sequence	Proposed features	K-nearest neighbours	0.00	66.3
		Random forest	0.03	66.6
		NaiveBayes	0.01	77.3
		Support vector machine	0.13	79.4
		Multilayer perceptron	11.63	83.3
Case 2: W = 10	Proposed features	NaiveBayes	0.20	77.0
		Multilayer perceptron	9.75	86.0
		Support vector machine	6.70	93.4
		Random forest	1.01	96.4
		K-nearest neighbours	0.02	99.8

- Case 1: we took all the distances between joints features and their velocities in a sliding window W = 3. We found that the Support Vector Machines (SVM) algorithms is the best classification algorithm with average accuracy 84.8 %, and it needed 65.4 s to complete the classification and validation.
- Case 2: We worked with our proposed features in Table 1 with a size of window equal to the number of frames in each scene. We found that multilayer perceptron

algorithm is the best algorithm, it achieves average accuracy 83.8 % and it needs 11.63 s.

- Case 3: We worked with our proposed features with a size of W = 10 for the sliding window and K-nearest neighbors for classification. Here we get the best result during all our work; we get an average accuracy 99.8 % and it needed 0.02 s.

5 Discussion

The main achievement in this work is that very high average accuracy of 99.8 % has been obtained using the proposed 10 features which are nine harmonic means of relational distances between pairs of six joints and one relational velocity between two joints, a sliding window of 10 frames and k-nearest neighbours classifier. For comparison, the authors of the dataset achieved 91.1 % [6] by using all geometric relational features based on distance between all pairs of joints. The number of these features is $435 \times W$ where W is the size of an extended sliding window which includes irrelevant frames around the main actions and they used Multiple Instance Learning in a boosting framework (MILBoost) to deal with the irrelevant frames in the training data. In our work, we noticed that there were no features related to the left hand or left foot of the actors. This is a drawback of the system since there were only right handed actors in the dataset.

6 Conclusion

We presented in this paper an approach for action recognition, violent actions included, using minimum number of features. These features are nine harmonic means of nine relational distances between nine pairs of selected six joints and one relational velocity between two joints. We choose these features depending on having the most useful attributes. At the end, k-nearest neighbours was employed for classification. Using harmonic mean concept shows very successful and easy method to deal with irrelevant frames to an action in a window. We used a general dataset of action as a benchmark since there is no complete dataset about violent actions using depth sensors so far. Our next step is to build our own dataset of violent actions and evaluate our methodology on our future dataset. Then, we will implement this approach in hardware as it uses small data for processing with simple algorithms.

Acknowledgement. The authors of this paper like to thank Yun *et al.* [6] for their generosity in sharing the dataset for use in this work.

References

1. Ke, S.R., Thuc, H.L.U., Lee, Y.J., Hwang, J.N., Yoo, J.H., Choi, K.H.: A review on video-based human activity recognition. Computers **2**(2), 88–131 (2013)
2. Li, W., Zhang, Z., Liu, Z.: Action recognition based on a bag of 3d points. In: Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE Computer Society Conference, pp. 9–14, California (2010)

3. Ni, B., Wang, G., Moulin, P.: Rgbd-hudaact: a color-depth video database for human daily activity recognition. Consumer Depth Cameras for Computer Vision, pp. 193–208. Springer, London (2013)
4. Masood, S.Z., Ellis, C., Nagaraja, A., Tappen, M. F., LaViola Jr., J.J., Sukthankar, R.: Measuring and reducing observational latency when recognizing actions. In: Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference, pp. 422–429. IEEE, Barcelona (2011)
5. Sung, J., Ponce, C., Selman, B., Saxena, A.: Human activity detection from RGBD images. Plan Act. Intent Recognit. **64**, 47–55 (2011)
6. Yun, K., Honorio, J., Chattopadhyay, D., Berg, T.L., Samaras, D.: Two-person interaction detection using body-pose features and multiple instance learning. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference, pp. 28–35. Rhode Island (2012)
7. Liu, T., Song, Y., Gu, Y., Li, A.: Human action recognition based on depth images from microsoft kinect. In: 2013 Fourth Global Congress on Intelligent Systems (GCIS), pp. 200–204. IEEE, Cape Town (2013)