

# Violence Classification Based on Shape Variations from Multiple Views

Fawang Liu and Yunde Jia

Beijing Laboratory of Intelligent Information Technology,  
School of Computer Science, Beijing Institute of Technology, Beijing 100081, PRC  
{liufawang, jiayunde}@bit.edu.cn

## Abstract

Most existing algorithms for human behavior analysis concentrate on action recognition through assuming that input sequences are well pre-segmented and restricting examples into a small vocabulary. In this paper, we present a novel action violence classification framework which directly evaluates the potential threat based on shape variations. We extract silhouettes as input features, employ the  $\mathfrak{R}$  transform to project binary shapes into the Radon space, and fuse multiple views to classify action violence. Experimental results on the INRIA IXMAS database demonstrate the efficiency and robustness of the proposed method.

## 1. Introduction

Visual surveillance is currently a hot research topic. Cameras are cheap and ubiquitous, but the manpower required to supervise them is expensive. Consequently, the video is often monitored sparingly or not at all. In fact, it is usually used merely as an archive, to refer back to once an incident has taken place. How to obtain a description of what is happening and take appropriate measures (for example alerting supervisors) becomes an emergent task of automatic surveillance.

Lots of papers on human behavior analysis have been published recently, and most approaches focus on action recognition through assuming that input sequences are well segmented and restricting the examples used into a small vocabulary [1]. In fact, we sometimes even feel it difficult to discriminate ambiguous actions such as “running” and “jogging”, and what most surveillance systems ultimately concern is the potential threat to oneself or others in the scene.

Generally, besides position changes, dangerous actions often imply intense pose variations (usually corresponding to violent shape deformations in 2D video sequences). In this paper, we dedicate to classify action

violence based on shape variations. In some situations, such as prisons or mental hospitals, the backgrounds are simple and persons are often required to wear uniform clothes in special colors, which makes our idea be potentially practical.

As shown in Fig.1, an integrated framework is proposed to classify action violence into three levels starting with blue, the lowest alert level, and followed by yellow and red. It avoids the requirements of action spotting and small vocabulary. Since shape is view-dependent, we integrate observations from multiple-views to improve the performance of evaluation. The  $\mathfrak{R}$  transform, which has a low time complexity and a nice behavior with respect to common geometrical transformations, is used to represent silhouettes.

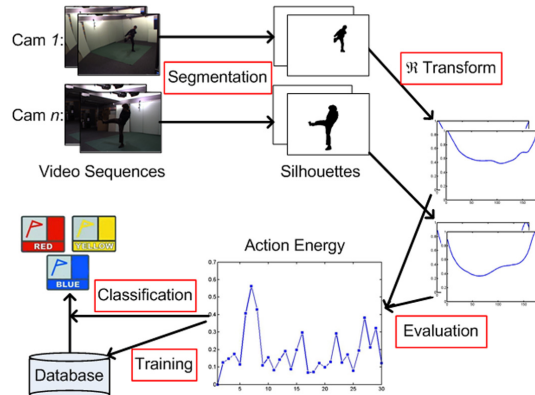


Figure 1. Flowchart of the framework.

## 2. Feature selection

Informative features are critical to activity analysis. Various techniques of feature representation have been explored in the past, and a brief review of shape-based feature descriptors is presented as follows.

Shape-based features are usually utilized because they can be extracted more efficiently and are more robust to appearance variations [2]. Contour and silhouette are two main classes of shape-based descriptors. The contour method only extracts the boundary of a shape, but the silhouette approach works on a shape as a whole taking into account the pixels within the shape.

Usual contour-based descriptors include Fourier descriptors, wavelets, and Hough transform [3][4]. Since contour descriptors are based on the shape boundary, they can not capture the internal structure and are not suited to disjoint shapes or shapes with holes. Consequently, they are limited to certain applications.

Common silhouette-based descriptors include invariant moment [5], Zernike moment [6], pseudo-Zernike moment [7], etc. Silhouette-based methods are very popular, but they are computational expensive and often need to normalize images to achieve common geometrical invariance. These normalization processes usually introduce errors and sensitivity to noise.

As a silhouette-based descriptor, the  $\mathfrak{R}$  transform has low computational cost and nice performance to common geometrical transformations [2]. In this work, we employ the  $\mathfrak{R}$  transform for feature representation.

### 3. Feature representation

The  $\mathfrak{R}$  transform is an improved representation of the Radon transform [8]. In mathematics, the 2D Radon transform is the transform consisting of the integral of a function over the lines taken at different angles. For a discrete binary image, each non-zero image point is projected to a Radon matrix. Let  $f(x, y)$  be an image. Its Radon transform is given by

$$T_{Rf}(\rho, \theta) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \delta(x, y) dx dy, \quad (1)$$

where  $\delta(x, y) = \delta(x \cos \theta + y \sin \theta - \rho)$ ,  $\rho \in (-\infty, \infty)$ ,  $\theta \in [0, \pi)$ , and  $\delta(\cdot)$  satisfies

$$\delta(x) = \begin{cases} 1 & \text{if } x = 0 \\ 0 & \text{otherwise} \end{cases}. \quad (2)$$

The Radon transform is sensitive to translation and scaling. To solve these problems, the  $\mathfrak{R}$  transform is introduced and defined as

$$\mathfrak{R}_f(\theta) = \int_{-\infty}^{\infty} T_{Rf}^2(\rho, \theta) d\rho. \quad (3)$$

The  $\mathfrak{R}$  transform is invariant under translation and scaling if it is normalized by a scaling factor.

### 4. Violence evaluation

Let  $\mathfrak{R}$  be the discrete  $\mathfrak{R}$  transform of a silhouette and  $\mathfrak{R}$  is a 180-dimensional vector, then the extracted feature can be represented as  $F(i) = \mathfrak{R}(i)$ , where  $i = 1, \dots, 180$ . Thus, the violence evaluation problem can be easily converted to assessing the resemblance of one curve with another.

The Hausdorff distance is adapted to measure the similarity between different shapes. Given  $F_{t-1}^k$  and  $F_t^k$ , the transform results of targets captured by camera  $k$  from successive pose  $P_{t-1}$  and  $P_t$ , the shape variation of point  $i$  is measured by

$$H^k(i) = \max(H_{P_{t-1}}^k(i), H_{P_t}^k(i)), \quad (4)$$

where

$$H_{P_{t-1}}^k(i) = \min_{-1 \leq c \leq 1} (\|F_{t-1}^k(i), F_t^k(i+c)\|), \quad (5)$$

$$H_{P_t}^k(i) = \min_{-1 \leq c \leq 1} (\|F_t^k(i), F_{t-1}^k(i+c)\|), \quad (6)$$

and  $\|\cdot\|$  is some underlying norm (e.g. the Euclidean norm or  $L_2$ ).  $H_{P_{t-1}}^k(i)$  measures the distance from point  $i$  of  $F_{t-1}^k$  to its corresponding neighbor points in  $F_t^k$  and takes the shortest distance as the result. The distance  $H^k(i)$  is the maximum of  $H_{P_{t-1}}^k(i)$  and  $H_{P_t}^k(i)$ . In another word,  $H^k(i)$  is the degree of mismatch by taking the neighbor points into account.

The result of violence evaluation from  $P_{t-1}$  to  $P_t$  observed by view  $k$  is

$$Egy^k(P_{t-1}, P_t) = \sum_{i=1}^{180} H^k(i). \quad (7)$$

To improve the robustness of violence evaluation, multi-view fusion is employed, and the final result is expressed by

$$Egy(P_{t-1}, P_t) = \max_{1 \leq k \leq m} (Egy^k(P_{t-1}, P_t)), \quad (8)$$

where  $m$  is the number of cameras.

### 5. Violence classification

To describe action violence more intuitively, we classify it into three levels: blue alert, yellow alert and red alert. Since defective segmentation will also lead to shape variation, a statistical method is explored.

To treat both isolated actions and continuous actions in the same framework and to reduce the influence of imperfect segmentation, we use a sliding window to extract feature sequences, each having  $n$  frames with an

**Table 1. Comparisons of average computing time (ACT).**

	IM	ZM	p-ZM	$\mathfrak{R}$ transform
ACT	0.042s	0.098s	0.137s	0.065s

overlap of  $\frac{n}{2}$  frames between the consecutive ones. We divide energy into  $l$  parts and calculate the number of energy points,  $N_1, \dots, N_l$ , in every part. The statistical result is defined as

$$V(\cdot) = \sum_{i=1}^l \alpha_i N_i, \quad (9)$$

where  $\alpha_i$  is the weight of  $N_i$ ,  $\alpha_1 < \alpha_2 < \dots < \alpha_l$ .

During the training process, a series of representative sequences are selected and divided into three categories to obtain two thresholds  $T_1$  and  $T_2$ . The action violence is determined by

$$level(\cdot) = \begin{cases} \text{red} & \text{if } V(\cdot) \geq T_2 \\ \text{yellow} & \text{if } V(\cdot) \in [T_1, T_2) \\ \text{blue} & \text{if } V(\cdot) < T_1 \end{cases} \quad (10)$$

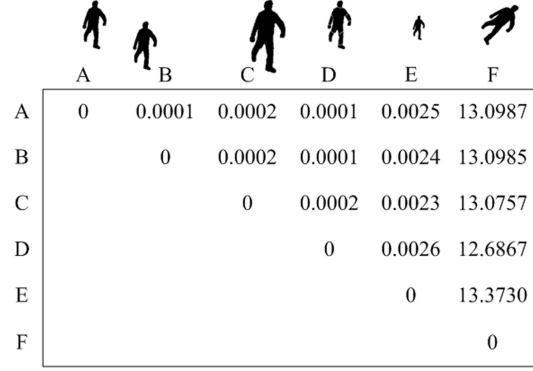
## 6. Experimental results

The multiple-video data used here are from INRIA Rhône-Alpes' multiple-camera platform Grimage and PERCEPTION research group. The database is available at <https://charibdis.inrialpes.fr>. It contains 13 daily-live actions performed by 11 actors, each 3 times, and viewed by 5 cameras. The actors choose freely position and orientation.

### 6.1. Performance of feature description

Since contour descriptors are not suitable for disjoint shapes or shapes with holes, only silhouette-based approaches are concerned here. For images of  $390 \times 291$  pixels, Table 1 lists the average computing time of invariant moment(IM), Zernike Moment (ZM), pseudo-Zernike (p-ZM) and the  $\mathfrak{R}$  transform. The results are obtained by MATLAB on a Pentium(R) Dual-Core 3.4GHz PC with 1G memory. The processing time of the  $\mathfrak{R}$  transform is just higher than that of IM.

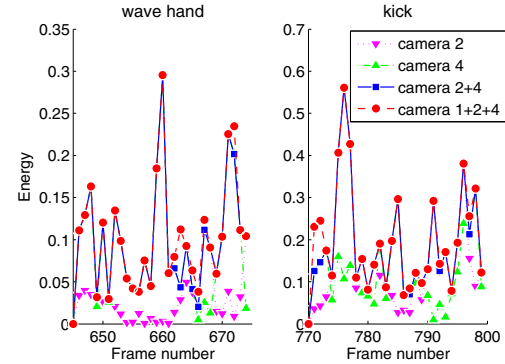
Fig.2 summarizes the results between pairs of variations for the joint shape A. From B to F, they are respectively translated randomly, increased in scale by 150%, manually segmented to 9 parts, scaled down by 50%, and rotated by  $45^\circ$ . All transform results are normalized by the shape area. Experimental results show that rotation will evidently affect the energy, but the influence of other variations is limited.



**Figure 2. Energy of shape variations.**

### 6.2. Performance of violence evaluation

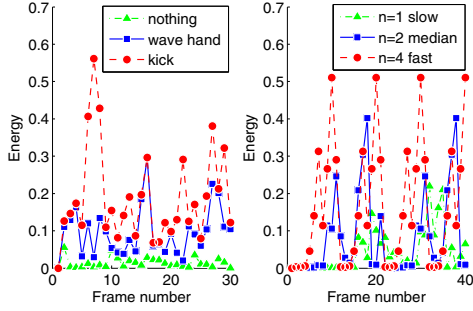
To improve the robustness of description, a multiple-view combination method is used. Fig.3 shows the estimation result of “wave hand” and “kick” based on single-view observation and multiple-view fusion. We can see that the combination of camera 2 and camera 4 is useful for violence description, but further adding cameras does not work much anymore. That is because those two cameras are approximately fronto-parallel and perpendicular one another and naturally they can describe the primary deformations.



**Figure 3. Single-view vs. multiple-view.**

Fig.4 depicts the energy comparisons of different actions and the same action at different motion speeds. To simulate the energy variations at different speeds, we extract one frame out of every  $n$  frames from the action “punch” performed by pao under camera 4 and duplicate the extracted frames.

As shown in Fig.4, “kick” is more violent than “wave hand” and “nothing”, and fast motion is more intense than slow motion for the same action.



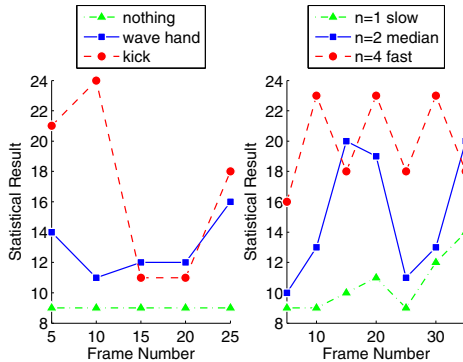
**Figure 4. Violence comparisons of different actions and the same action at different motion speeds.**

### 6.3. Performance of violence classification

Based on the observation that action energy mainly resides in  $[0, 1]$ ,  $n$  and  $l$  are empirically set to 9 and 10, and  $Egy(\cdot)$  is divided and labeled according to

$$i = \begin{cases} 10 & \text{if } Egy(\cdot) \geq 0.9 \\ \lfloor Egy(\cdot) \times 10 \rfloor + 1 & \text{otherwise} \end{cases}, \quad (11)$$

where  $\lfloor x \rfloor$  gets the nearest integer less than or equal to  $x$ ,  $i \in [1, l]$ . The weight of  $N_i$  is defined as  $\alpha_i = i$ . Fig.5 shows the statistical result of actions in Fig. 4.



**Figure 5. Statistical result of action violence in Fig. 4.**

To confirm the classification performance, 100 clips (about 4000 frames) are selected and 40 sequences are employed for training. The thresholds,  $T_1$  and  $T_2$ , are fixed according to the means and variances of the training examples. Different descriptors are employed respectively for feature representation. Table 2 shows

**Table 2. Comparisons of correct classification rates (CCR).**

	IM	ZM	p-ZM	$\mathfrak{R}$ transform
CCR	79.8%	71.6%	73.2%	91.3%

the correct classification rates compared with the hand-marked ground truth.

## 7. Conclusion

In this paper, we have presented a general framework for classifying action violence based on shape variations from multiple views. Our method does not require video alignment and is more robust to disjoint shapes due to the adoption of the  $\mathfrak{R}$  transform. Furthermore, the proposed approach avoids the requirements of action spotting and small vocabulary. The future work will focus on the evaluation of our classification framework on larger datasets.

## Acknowledgement

This work was partially supported by the Natural Science Foundation of China (60675021) and the Chinese High-Tech Program (2006AA01Z120).

## References

- [1] T. B. Moeslund, A. Hilton, and V. Kruger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(1): 90-126, 2006.
- [2] S. Tabbone, L. Wendling, and J. Salmon. A new shape descriptor defined on the Radon Transform. *Computer Vision and Image Understanding*, 102(1): 42-51, 2006.
- [3] D. Zhang and G. Lu. Study and evaluation of different Fourier methods for image retrieval. *Image and Vision Computing*, 23(1): 33-49, 2005.
- [4] C. Chuang and C. Kuo. Wavelet Descriptor of Planar Curves: Theory and Applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(1), 1999.
- [5] M. Hu. Visual pattern recognition by Moment invariants. *IRE Transactions on Information Theory*, 8(1): 179-187, 1962.
- [6] M. R. Teague. Image analysis via the general theory of moments. *Journal of Optimal Society of American*, 70(8): 920-930, 1980.
- [7] R. Mukundan and K. R. Ramakrishnan. *Moment Functions in Image Analysis: Theory and Applications*. World Scientific Publishing Singapore, 1998.
- [8] S. R. Deans. *Applications of the Radon Transform*. Wiley Inter-science Publications, New York, 1983.