# Fast Violence Detection in Video

Oscar Deniz[1], Ismael Serrano[1], Gloria Bueno[1] and Tae-Kyun Kim[2]
[1]*VISILAB group, University of Castilla-La Mancha, E.T.S.I.Industriales, Avda. Camilo Jose Cela s/n, Ciudad Real, 13071 Spain*
[2]*Department of Electrical and Electronic Engineering, Imperial College, South Kensington Campus, London SW7 2AZ, U.K.*

Keywords: Action Recognition, Violence Detection, Fight Detection.

Abstract: Whereas the action recognition problem has become a hot topic within computer vision, the detection of fights or in general aggressive behavior has been comparatively less studied. Such capability may be extremely useful in some video surveillance scenarios like in prisons, psychiatric centers or even embedded in camera phones. Recent work has considered the well-known Bag-of-Words framework often used in generic action recognition for the specific problem of fight detection. Under this framework, spatio-temporal features are extracted from the video sequences and used for classification. Despite encouraging results in which near 90% accuracy rates were achieved for this specific task, the computational cost of extracting such features is prohibitive for practical applications, particularly in surveillance and media rating systems. The task of violence detection may have, however, specific features that can be leveraged. Inspired by psychology results that suggest that kinematic features alone are discriminant for specific actions, this work proposes a novel method which uses extreme acceleration patterns as the main feature. These extreme accelerations are efficiently estimated by applying the Radon transform to the power spectrum of consecutive frames. Experiments show that accuracy improvements of up to 12% are achieved with respect to state-of-the-art generic action recognition methods. Most importantly, the proposed method is at least 15 times faster.

## 1 INTRODUCTION

In the last years, the problem of human action recognition from video has become tractable by using computer vision techniques, see for example the survey (Poppe, 2010). Despite its potential usefulness, the specific task of violent action detection has been comparatively less studied. A violence detector has, however, immediate applicability in the surveillance domain. The primary function of large-scale surveillance systems deployed in institutions such as schools, prisons and psychiatric care facilities is for alerting authorities to potentially dangerous situations. However, human operators are overwhelmed with the number of camera feeds and manual response times are slow, resulting in a strong demand for automated alert systems. Similarly, there is increasing demand for automated rating and tagging systems that can process the great quantities of video uploaded to websites. Violence detection is becoming important not only on an application level but also on a more scientific level, because it has particularities that make it different from generic action recognition. For all these reasons the interest in violence detection has been steadily growing, and different proposals are already being published in major journals and conferences. Also, public datasets are becoming increasingly available that are specifically designed for this task.

One of the first proposals for violence recognition in video is Nam *et al.* (Nam et al., 1998), which proposed recognizing violent scenes in videos using flame and blood detection and capturing the degree of motion, as well as the characteristic sounds of violent events. Cheng *et al.* (Cheng et al., 2003) recognizes gunshots, explosions and car-braking in audio using a hierarchical approach based on Gaussian mixture models and Hidden Markov models (HMM). Giannakopoulos *et al.* (Giannakopoulos et al., 2006) also propose a violence detector based on audio features. Clarin *et al.* (Clarin et al., 2005) present a system that uses a Kohonen self-organizing map to detect skin and blood pixels in each frame and motion intensity analysis to detect violent actions involving blood. Zajdel *et al.* (Zajdel et al., 2007), introduced the CASSANDRA system, which employs motion features re-


Deniz O., Serrano I., Bueno G. and Kim T..
Fast Violence Detection in Video.
DOI: 10.5220/0004695104780485
In *Proceedings of the 9th International Conference on Computer Vision Theory and Applications* (VISAPP-2014), pages 478-485
ISBN: 978-989-758-004-8

lated to articulation in video and scream-like cues in audio to detect aggression in surveillance videos.

More recently, Gong *et al.* (Gong et al., 2008) propose a violence detector using low-level visual and auditory features and high-level audio effects identifying potential violent content in movies. Chen *et al.* (Chen et al., 2008) use binary local motion descriptors (spatio-temporal video cubes) and a bag-of-words approach to detect aggressive behaviors. Lin and Wang (Lin and Wang, 2009) describe a weakly-supervised audio violence classifier combined using co-training with a motion, explosion and blood video classifier to detect violent scenes in movies. Giannakopoulos *et al.* (Giannakopoulos et al., 2010) present a method for violence detection in movies based on audio-visual information that uses a statistics of audio features and average motion and motion orientation variance features in video combined in a k-Nearest Neighbor classifier to decide whether the given sequence is violent. Chen *et al.* (Chen et al., 2011) proposed a method based on motion and detecting faces and nearby blood. Violence detection has been even approached using static images (Wang et al., 2012). Also recently, (Zou et al., 2012) approached the problem within the context of video sharing sites by using textual tags along with audio and video. Proof of the growing interest is also the MediaEval Affect Task, a competition that aims at discovering violence in color movies (Demarty et al., 2012). In this case the algorithms have access to additional information such as audio, subtitles and previously-annotated concepts. Besides, no comparisons are made about computational times.

In summary, a number of previous works require audio cues for detecting violence or rely on color to detect cues such as blood. In this respect, we note that there are important applications, particularly in surveillance, where audio and color are not available. Besides, while explosions, blood and running may be useful cues for violence in action movies, they are rare in real-world situations. In any case, violence detection *per se* is an extremely difficult problem, since violence is a subjective concept. Fight detection, on the contrary, is a specific violence-related task that may be tackled using action recognition techniques and which has immediate applications.

Whereas there is a number of well-studied datasets for action recognition, significant datasets with violent actions (fights) have not been made available until the work (Bermejo et al., 2011). In that work the authors demonstrated encouraging results on violence detection, achieving 90% accuracy using MoSIFT features ((Chen et al., 2010)). MoSIFT descriptors are obtained from salient points in two parts: the first is an aggregated histogram of gradients (HoG) which describe the spatial appearance. The second part is an aggregated histogram of optical flow (HoF) which indicates the movement of the feature point. Despite being considered within state-of-the-art action recognition methods, the computational cost of extracting these features is prohibitively large, taking near 1 second per frame on a high-end laptop. This precludes use in practical applications, where many camera streams may have to be processed in real-time. Such cost is also a major problem when the objective is to embed a fight detection functionality into a smart camera (i.e. going from the extant embedded motion detection to embedded violent motion detection).

Features such as MoSIFT encode both motion and appearance information. However, research on human perception of other's actions (using point-light displays, see Figure 1) has shown that the kinematic pattern of movement is sufficient for the perception of actions (Blake and Shiffrar, 2007). This same idea has been also supported by research on the computer vision side (Oshin et al., 2011; Bobick and Davis, 1996). More specifically, empirical studies in the field have shown that relatively simple dynamic features such as velocity and acceleration correlate to emotional attributes perceived from the observed actions (Saerbeck and Bartneck, 2010; Clarke et al., 2005; Castellano et al., 2007; Hidaka, 2012), albeit the degree of correlation varies for different emotions. Thus, features such as acceleration and jerkiness tend to be associated to emotions with high activation (eg. anger, happiness), whereas slow and smooth movements are more likely to be judged as emotions with low activation (eg. sadness).

In this context, this work assumes that fights in video can be reliably detected by such kinematic cues that represent violent motion and strokes. Since extreme accelerations play a key role we propose a novel method to infer them in an efficient way. The proposed fight detector attains better accuracy rates than state-of-the-art action recognition methods at much less computational cost. The paper is organized as follows. Section 2 describes the proposed method. Section 3 provides experimental results. Finally, in Section 4 the main conclusions are outlined.

## 2 PROPOSED METHOD

As mentioned above, the presence of large accelerations is key in the task of violence recognition. In this context, body part tracking can be considered, as in (Datta et al., 2002), which introduced the so-

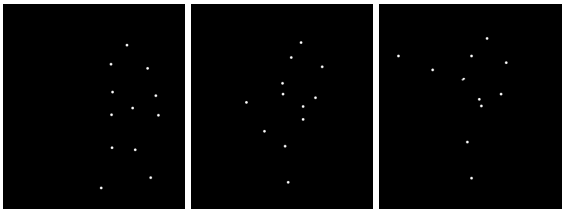

Figure 1: Three frames in a point light display movie depicting a karate kick.

called Acceleration Measure Vectors (AMV) for violence detection. In general, acceleration can be inferred from tracked point trajectories. However, we have to note that extreme acceleration implies image blur (see for example Figure 2), which makes tracking less precise or even impossible.

Motion blur entails a shift in image content towards low frequencies. Such behavior allows to build an efficient acceleration estimator for video. First, we compute the power spectrum of two consecutive frames. It can be shown that, when there is a sudden motion between the two frames, the power spectrum image of the second frame will depict an ellipse (Barlow and Olshausen, 2004). The orientation of the ellipse is perpendicular to the motion direction, the frequencies outside the ellipse being attenuated, see Figure 3. Most importantly, the eccentricity of this ellipse is dependent on the acceleration. Basically, the proposed method aims at detecting the sudden presence of such ellipse. In the following, the method is described in detail.

Figure 2: Two consecutive frames in a fight clip from a movie. Note the blur on the left side of the second frame.

Let $I_{i-1}$ and $I_i$ be two consecutive frames. Motion

Figure 3: Left: Sample image. Center: simulated camera motion at 45°. Right: Fourier transform of the center image.

blur is equivalent to applying a low-pass oriented filter $C$.

$$\mathcal{F}(I_i) = \mathcal{F}(I_{i-1}) \cdot C \tag{1}$$

where $\mathcal{F}(\cdot)$ denotes the Fourier Transform. Then:

$$C = \frac{\mathcal{F}(I_i)}{\mathcal{F}(I_{i-1})} \tag{2}$$

The low-pass oriented filter in $C$ is the above-mentioned ellipse.

For each pair of consecutive frames, we compute the power spectrum using the 2D Fast Fourier Transform (in order to avoid edge effects, a Hanning window was applied before computing the FFT). Let us call these spectra images $P_{i-1}$ and $P_i$. Next, we simply compute the image:

$$C = \frac{P_i}{P_{i-1}} \tag{3}$$

When there is no change between the two frames, the power spectra will be equal and $C$ will have a constant value. When motion has occurred, an ellipse will appear in $C$. Our objective is then to detect such ellipse and estimate its eccentricity, which represents the magnitude of the acceleration. Ellipse detection can be reliably performed using the Radon transform, which provides image projections along lines with different orientations, see Figure 4.

After computing the Radon transform image $R$, its vertical *maximum* projection vector $vp$ is obtained and normalized to maximum value 1 (see Figure 4-bottom). When there is an ellipse in $C$, this vector will show a sharp peak, representing the major axis of the ellipse. The kurtosis $K$ of this vector is therefore taken as an estimation of the acceleration.

Note that kurtosis alone cannot be used as a measure, since it is obtained from a normalized vector (i.e. it is dimensionless). Thus, the average power per pixel $P$ of image $C$ is also computed and taken as an additional feature. Without it, any two frames could lead to high kurtosis even without significant motion.

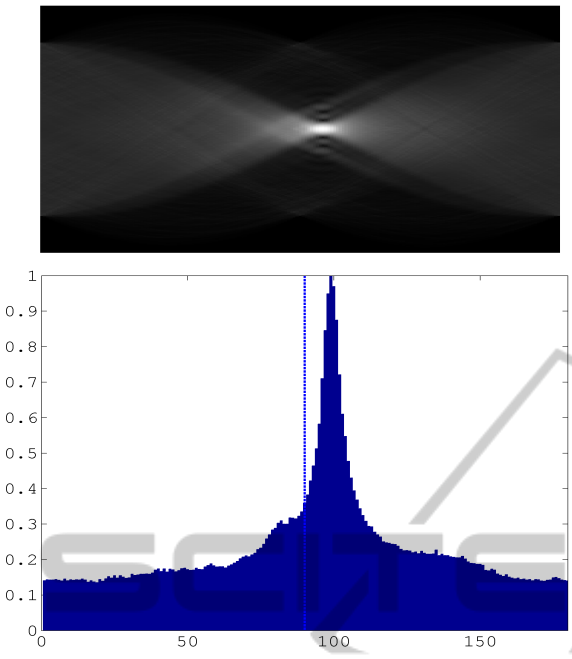


Figure 4: Top: Radon transform image of Figure 3-left under a simulated camera motion at 98°. The horizontal axis represents angles between 0 and 180°. Bottom: vertical projection of the Radon image.

The previous paragraphs have described a procedure that obtains two features $K$ and $P$ for each pair of consecutive frames. Deceleration was also considered as an additional feature, and it can be obtained by swapping the consecutive frames and applying the same algorithm explained above. For video sequences, we compute histograms of these features, so that acceleration/deceleration patterns can be inferred.

In a variant of the proposed method, ellipse eccentricity can be estimated by first locating the position $p$ of the maximum of $vp$. This maximum is associated to the major axis of the ellipse. The minor axis is then located at position:

$$\begin{aligned} q &= p+90^\circ \quad if(p+90^\circ) \leq 180^\circ \\ q &= p-90^\circ \quad otherwise \end{aligned}$$

The ratio of the two values may then be used as a feature, instead of the kurtosis:

$$r = \frac{vp(p)}{vp(q)} \tag{4}$$

Algorithm 1 shows the detailed steps of the proposed method.

Since the proposed method does not involve tracking or optical-flow techniques it is more suitable for measuring extreme accelerations. Lastly, it is important to note that global (i.e. camera) motion could also cause blur in the image. In order to remove such blur, it is necessary to perform a deconvolution preprocessing step. The phase correlation technique is first used to infer global motion between each pair of consecutive frames. If global motion is detected, the estimated angle and length of the displacement is used to form a PSF with which to perform deconvolution of the second frame (we used the Lucy-Richardson iterative deconvolution method). This is intended to remove the blur caused by global motion (camera motion), while any local blurs will remain. The method described above is then applied to the pair of frames as shown in Algorithm 1 above.

When backgrounds are relatively uniform and displacements small, global motion estimation may still fail. The fail mode is typically represented by real global motion which goes undetected, i.e. an incorrect (0,0) displacement is estimated. Since the proposed method is heavily dependent on global motion,

**Input**: $S$ = (Short) sequence of gray scale images. Each image in $S$ is denoted as $f_{x,y,t}$, where $x = 1,2,\ldots,N$, $y = 1,2,\ldots,M$ and $t = 1,2,\ldots,T$.
**Result**: $3 \cdot n\_bins$ discriminant features

**for** $t = 1$ *to* $T$ **do**

1. Apply a Hanning Window to $f_{x,y,t}$:

$$g_{x,y,t} = f_{x,y,t} \cdot H_{x,y}$$

where $H_{x,y} = h(N) \cdot h(M)'$ and $h$ is a column vector given by:

$$h(L) = \tfrac{1}{2}\left[1 - \cos\left(2\pi \tfrac{l}{L}\right)\right], \text{ for } l = 1,2,\cdots,L$$

2. Apply FFT to $f_{x,y,t}$: $F_{v,w,t} = \mathcal{F}(g_{x,y,t})$
3. Compute $C_{v,w} = F_{v,w,t}/F_{v,w,t-1}$
4. Compute Radon transform of $C$: $R_{d,\theta} = \mathcal{R}(C_{v,w})$
5. Compute vertical max projection of $R$: $p_\theta = \max_d(R_{d,\theta})$
6. Normalize $p_\theta = p_\theta/m$, where $m = \max_\theta(p_\theta)$
7. Compute feature $A_t = Kurtosis(p_\theta)$
8. Compute feature $P_t = mean_{v,w}C_{v,w}$
9. Compute feature $D_t$ (deceleration) using the same steps above but swapping $t$ for $t-1$

**end**
**return** $Histogram(A,n\_bins), Histogram(P,n\_bins), Histogram(D,n\_bins)$

**Algorithm 1:** Algorithm for computing the main features in the proposed method.

further measures must be taken in practice to at least detect the presence of global motion *versus* local motion. The window function mentioned above restricts processing to the inner part of the image. It is reasonable to assume that, when motion is global, changes in the outer part of the image will be relatively on par with those in the inner part. Thus, an additional 'Outer Energy' $O$ feature was computed and used in the same way as the others:

$$O_{x,y,t} = \frac{|f_{x,y,t} - f_{x,y,t-1}| \cdot (1 - H_{x,y})}{M \cdot N} \quad (5)$$

The mean and standard deviation of $O$ are then used as additional features.

## 3 EXPERIMENTS

The work (Bermejo et al., 2011) introduced the first two datasets explicitly designed for assessing fight detection. The first dataset ("Hockey") consists of 1000 clips at a resolution of 720x576 pixels, divided in two groups, 500 fights (see Fig. 5 top) and 500 non-fights, extracted from hockey games of the National Hockey League (NHL). Each clip was limited to 50 frames and resolution lowered to 320x240. The second dataset ("Movies") introduced in (Bermejo et al., 2011) consists of 200 video clips in which fights were extracted from action movies (see Figure 5 bottom). The non-fight videos were extracted from public action recognition datasets. Unlike the hockey dataset, which was relatively uniform both in format and content, these videos depicted a wider variety of scenes and were captured at different resolutions.

In the experiments, the Radon transform was computed between 0 and 180 in steps of $\theta = 20$ degrees. 4-bin histograms were computed for each of the three main features (acceleration, deceleration and power, see the previous Section). The results measured using 10-fold cross-validation are shown in Table 1. For convenience we also show the results reported in (Bermejo et al., 2011), which used an SVM classifier.

In (Bermejo et al., 2011) STIP features performed poorly on the Movie dataset and so MoSIFT was considered the best descriptor. MoSIFT's superiority has been also proven in other action recognition works. The proposed method gives roughly equivalent accuracy and AUC for the Hockey dataset whereas it improves on the Movie dataset by 9%.

Since the proposed method is based on extreme acceleration patterns, energetic actions may pose a problem. However, the method performs quite well in this respect, as evidenced in the Hockey dataset results. Although the Hockey dataset may represent the most difficult dataset for a fight detector, in practice we aim at separating fights from other actions. Consequently, a more challenging dataset was also considered. The UCF101 (Soomro et al., 2012) is a data set of realistic action videos collected from YouTube, having 101 action categories. UCF101, see Figure 6, gives the largest diversity in terms of actions and with the presence of large variations in camera motion, object appearance and pose, object scale, viewpoint, cluttered background and illumination conditions it is the most challenging dataset to date. For our case, it is even more challenging since it includes 50 actions from sports. To our knowledge, this is the largest and most challenging dataset in which a fight detection algorithm has been tested.

Figure 5: Sample fight videos from the Hockey (top) dataset and the action movie (bottom) dataset.

In the experiments with UCF101, for the fight set we pooled the fight clips of both the Hockey and Movies dataset plus two of the 101 UCF actions that actually represented fights ("Punching" and "Sumo"). This gave a total of 1843 fight clips. Non-fight clips were taken from the other 99 action categories (42278 non-fight clips, totaling approximately 2 Million frames). In order to avoid unbalanced sets we used randomly chosen subsets of 500 fight and 500 non-fight clips. For each subset we performed a 10-fold cross-validation. This was in turn repeated 10 times.

Table 1: Results on the Hockey and Movies datasets, 5 runs of 10-fold cross-validation. Note: A stands for accuracy, DR stands for detection rate, FR stands for false positive rate, AUC stands for area under the (ROC) curve. In bold are shown the best accuracies for each dataset and classifier.

| Features | Classifier | Measure | Dataset | |
|---|---|---|---|---|
| | | | Movies | Hockey |
| BoW(STIP) | SVM | A | 82.5 ± 1.12 | 88.6 ± 0.15 |
| | | DR | 83.4 ±1.14 | 93.6 ±0.22 |
| | | FR | 18.4 ±1.14 | 16.5 ±0.18 |
| | | AUC | 0.8844 | 0.9383 |
| | Adaboost | A | 74.3 ±2.31 | 86.5 ±0.19 |
| | | DR | 70.2 ±3.70 | 88.7 ±0.76 |
| | | FR | 21.6 ±3.21 | 15.7 ±0.52 |
| | | AUC | 0.8121 | 0.9220 |
| BoW(MoSIFT) | SVM | A | 84.2 ±1.15 | **91.2** ±0.24 |
| | | DR | 100 ±0 | 92 ±0.17 |
| | | FR | 31.6 ±2.30 | 9.6 ±0.41 |
| | | AUC | 0.9267 | 0.9547 |
| | Adaboost | A | 86.5 ±1.58 | 89.5 ±0.40 |
| | | DR | 99.6 ±0.55 | 90.1 ±0.88 |
| | | FR | 26.6 ±3.05 | 11.1 ±0.27 |
| | | AUC | 0.9518 | 0.9492 |
| Proposed | SVM | A | **85.4** ±9.33 | 90.1 ±0 |
| | | DR | 71.4 ±19.42 | 80.2 ±0 |
| | | FR | 0.8 ±0.83 | 0 ±0 |
| | | AUC | 0.7422 | 0.9480 |
| | Adaboost | A | **98.9** ±0.22 | **90.1** ±0 |
| | | DR | 97.8 ±0.45 | 80.2 ±0 |
| | | FR | 0.0 ±0.0 | 0 ±0 |
| | | AUC | 0.9999 | 0.9020 |

Table 2: Results on the UCF101 dataset. Note: A stands for accuracy, DR stands for detection rate, FR stands for false positive rate, AUC stands for area under the (ROC) curve. In bold are shown the best accuracies for each classifier.

| Features | Classifier | Measure | Dataset |
|---|---|---|---|
| BoW(STIP) | SVM | A | 72 ±1.78 |
| | | DR | 86.2 ±1.83 |
| | | FR | 42.2 ±3.26 |
| | | AUC | 0.7352 |
| | Adaboost | A | 63.4 ±2.39 |
| | | DR | 75.3 ±3.60 |
| | | FR | 48.5 ±4.40 |
| | | AUC | 0.6671 |
| BoW(MoSIFT) | SVM | A | 81.3 ±0.78 |
| | | DR | 90.8 ±1.34 |
| | | FR | 28.1 ±1.88 |
| | | AUC | 0.8715 |
| | Adaboost | A | 51.3 ±0.32 |
| | | DR | 100 ±0 |
| | | FR | 97.4 ±0.64 |
| | | AUC | 0.5340 |
| Proposed | SVM | A | **93.4** ±6.09 |
| | | DR | 87.3 ±11.12 |
| | | FR | 0.45 ±1.28 |
| | | AUC | 0.9439 |
| | Adaboost | A | **92.8** ±6.29 |
| | | DR | 85.7 ±12.53 |
| | | FR | 0.02 ±0.06 |
| | | AUC | 0.9379 |

For BoW(MoSIFT), and even with the use of parallel K-means, extracting vocabularies from the whole dataset was unfeasible. Therefore, a random subset of samples was first selected (600 of each class) and then a vocabulary of size 500 (the best vocabulary size in (Bermejo et al., 2011)) was computed. The results are shown on Table 2. Figure 7 shows the ROC curve obtained for both methods with the SVM classifier. These results suggest that the method may effectively work as a fight detector for generic settings. Global motion estimation experiments did not seem to improve results significantly in this case either.

Note that the results show the higher detection rate already hypothesized in Section 2. This is evidenced by the resulting ROC curves closer to the vertical axis for the proposed method.

Table 3 shows the number of features used for classification and the computational cost measured (for feature extraction). The code for both STIP and MoSIFT was compiled. The code for the proposed method was interpreted and used no parallelization. These results show an improvement in speed of roughly 15 times with respect to the best previous method (MoSIFT). The fact that only 14 features are necessary (MoSIFT used 500) is an additional advantage for practical implementations.

Table 3: Feature extraction times. Average times measured with the non-fight videos in the UCF101 dataset, on an Intel Xeon computer with 2 processors at 2.90Ghz.

| **Method** | **Secs/frame** |
|---|---|
| MoSIFT | 0.6615 |
| STIP | 0.2935 |
| Proposed Method | 0.0419 |

# 4 CONCLUSIONS

Based on the observation that kinematic information may suffice for human perception of other's actions, in this work a novel detection method is proposed which uses extreme acceleration patterns as the main discriminating feature. The method shows promising features for surveillance scenarios and it also performs relatively well when considering challenging actions such as those that occur in sports. Accuracy improvements of up to 12% with respect to state-of-the-art generic action recognition techniques were achieved. We hypothesize that when motion is sufficient for recognition, appearance not only takes significant additional computation but it also may confuse the detector. Another interpretation is that a sort of overfitting may be occurring in that case. In any case, the extreme acceleration estimation proposed seems to perform well, given that other methods may fail because of the associated image blur.

The proposed method makes no assumptions on number of individuals (it can be also used to detect vandalism), body part detection or salient point track-

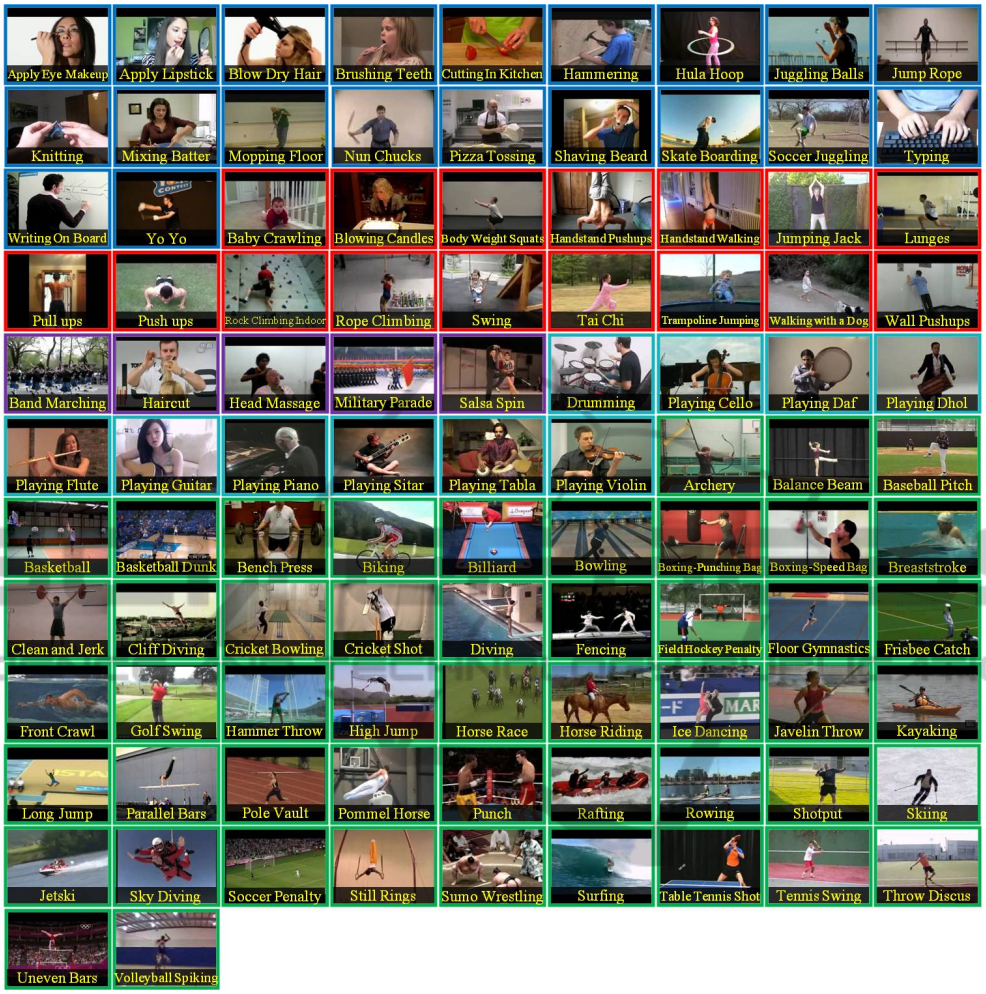


Figure 6: The 101 actions in UCF101 shown with one sample frame.

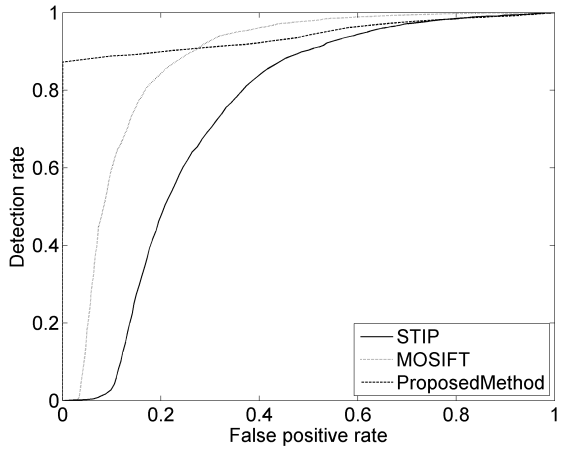


Figure 7: ROC curve with the SVM classifier. Average of 10 experimental runs.

ing. Besides, it is at least 15 times faster and uses only 14 features, which opens up the possibility of practical implementations. When maximum accuracy is needed, the method could also act as a first attentional stage in a cascade framework that also uses STIP or MoSIFT features.

Future work will seek to perfect the method by approximating the Radon transform, which is the most time-consuming stage. On a more basic level, we shall investigate the implications with regards to the relative importance of motion and appearance information for the recognition of certain actions.

## ACKNOWLEDGEMENTS

This work has been supported by research project TIN2011-24367 from the Spanish Ministry of Economy and and Competitiveness.

# REFERENCES


Barlow, H. B. and Olshausen, B. A. (2004). Convergent evidence for the visual analysis of optic flow through anisotropic attenuation of high spatial frequencies. *Journal of Vision*, 4(6):415–426.

Bermejo, E., Deniz, O., Bueno, G., and Sukthankar, R. (2011). Violence detection in video using computer vision techniques. In *14th Int. Congress on Computer Analysis of Images and Patterns*, pages 332–339.

Blake, R. and Shiffrar, M. (2007). Perception of Human Motion. *Annual Review of Psychology*, 58(1):47–73.

Bobick, A. and Davis, J. (1996). An appearance-based representation of action. In *Pattern Recognition, 1996., Proceedings of the 13th International Conference on*, volume 1, pages 307–312 vol.1.

Castellano, G., Villalba, S., and Camurri, A. (2007). Recognising human emotions from body movement and gesture dynamics. In Paiva, A., Prada, R., and Picard, R., editors, *Affective Computing and Intelligent Interaction*, volume 4738 of *Lecture Notes in Computer Science*, pages 71–82. Springer Berlin Heidelberg.

Chen, D., Wactlar, H., Chen, M., Gao, C., Bharucha, A., and Hauptmann, A. (2008). Recognition of aggressive human behavior using binary local motion descriptors. In *Engineering in Medicine and Biology Society*, pages 5238–5241.

Chen, L.-H., Su, C.-W., and Hsu, H.-W. (2011). Violent scene detection in movies. *IJPRAI*, 25(8):1161–1172.

Chen, M.-y., Mummert, L., Pillai, P., Hauptmann, A., and Sukthankar, R. (2010). Exploiting multi-level parallelism for low-latency activity recognition in streaming video. In *MMSys '10: Proceedings of the first annual ACM SIGMM conference on Multimedia systems*, pages 1–12, New York, NY, USA. ACM.

Cheng, W.-H., Chu, W.-T., and Wu, J.-L. (2003). Semantic context detection based on hierarchical audio models. In *Proceedings of the ACM SIGMM workshop on Multimedia information retrieval*, pages 109–115.

Clarin, C., Dionisio, J., Echavez, M., and Naval, P. C. (2005). DOVE: Detection of movie violence using motion intensity analysis on skin and blood. Technical report, University of the Philippines.

Clarke, T. J., Bradshaw, M. F., Field, D. T., Hampson, S. E., and Rose, D. (2005). The perception of emotion from body movement in point-light displays of interpersonal dialogue. *Perception*, 34:1171–1180.

Datta, A., Shah, M., and Lobo, N. D. V. (2002). Person-on-person violence detection in video data. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 1, pages 433–438.

Demarty, C., Penet, C., Gravier, G., and Soleymani, M. (2012). MediaEval 2012 affect task: Violent scenes detection in Hollywood movies. In *MediaEval 2012 Workshop Proceedings*, Pisa, Italy.

Giannakopoulos, T., Kosmopoulos, D., Aristidou, A., and Theodoridis, S. (2006). Violence content classification using audio features. In *Advances in Artificial Intelligence*, volume 3955 of *Lecture Notes in Computer Science*, pages 502–507.

Giannakopoulos, T., Makris, A., Kosmopoulos, D., Perantonis, S., and Theodoridis, S. (2010). Audio-visual fusion for detecting violent scenes in videos. In *6th Hellenic Conference on AI, SETN 2010, Athens, Greece, May 4-7, 2010. Proceedings*, pages 91–100, London, UK. Springer-Verlag.

Gong, Y., Wang, W., Jiang, S., Huang, Q., and Gao, W. (2008). Detecting violent scenes in movies by auditory and visual cues. In *Proceedings of the 9th Pacific Rim Conference on Multimedia*, pages 317–326, Berlin, Heidelberg. Springer-Verlag.

Hidaka, S. (2012). Identifying kinematic cues for action style recognition. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, pages 1679–1684.

Lin, J. and Wang, W. (2009). Weakly-supervised violence detection in movies with audio and video based co-training. In *Proceedings of the 10th Pacific Rim Conference on Multimedia*, pages 930–935, Berlin, Heidelberg. Springer-Verlag.

Nam, J., Alghoniemy, M., and Tewfik, A. (1998). Audio-visual content-based violent scene characterization. In *Proceedings of ICIP*, pages 353–357.

Oshin, O., Gilbert, A., and Bowden, R. (2011). Capturing the relative distribution of features for action recognition. In *Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 111–116.

Poppe, R. (2010). A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6):976 – 990.

Saerbeck, M. and Bartneck, C. (2010). Perception of affect elicited by robot motion. In *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction*, HRI '10, pages 53–60, Piscataway, NJ, USA. IEEE Press.

Soomro, K., Zamir, A., and Shah, M. (2012). UCF101: A dataset of 101 human action classes from videos in the wild. CRCV-TR-12-01. Technical report.

Wang, D., Zhang, Z., Wang, W., Wang, L., and Tan, T. (2012). Baseline results for violence detection in still images. In *AVSS*, pages 54–57.

Zajdel, W., Krijnders, J., Andringa, T., and Gavrila, D. (2007). CASSANDRA: audio-video sensor fusion for aggression detection. In *Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on*, pages 200–205.

Zou, X., Wu, O., Wang, Q., Hu, W., and Yang, J. (2012). Multi-modal based violent movies detection in video sharing sites. In *IScIDE*, pages 347–355.