Violence Detection in Video Using Computer Vision Techniques

Enrique Bermejo Nievas[1], Oscar Deniz Suarez[1], Gloria Bueno García[1], and Rahul Sukthankar[2]

[1] E.T.S.I.Industriales, Universidad de Castilla-La Mancha
Avda. Camilo Jose Cela s/n, 13071 Ciudad Real, Spain
Oscar.Deniz@uclm.es,
http://visilab.etsii.uclm.es/

[2] Intel Labs Pittsburgh and Robotics Institute, Carnegie Mellon, USA

Abstract. Whereas the action recognition community has focused mostly on detecting simple actions like clapping, walking or jogging, the detection of fights or in general aggressive behaviors has been comparatively less studied. Such capability may be extremely useful in some video surveillance scenarios like in prisons, psychiatric or elderly centers or even in camera phones. After an analysis of previous approaches we test the well-known Bag-of-Words framework used for action recognition in the specific problem of fight detection, along with two of the best action descriptors currently available: STIP and MoSIFT. For the purpose of evaluation and to foster research on violence detection in video we introduce a new video database containing 1000 sequences divided in two groups: fights and non-fights. Experiments on this database and another one with fights from action movies show that fights can be detected with near 90% accuracy.

Keywords: action recognition, fight detection, video surveillance

1 Introduction

In the last years, the problem of human action recognition at a distance has become tractable by using computer vision techniques. Although the first approaches obtained good results, they have some limitations too. There are, for example, aperture problems and discontinuities in optical flow based approaches [8], and illumination and reinitialization problems in feature tracking approaches [2]. More recently, the use of feature descriptors around interest points has become popular within the action recognition community, see the recent survey [16]. This approach analyzes actions by considering the video sequence as a space-time volume and using gradients, intensities, flows or other local features. This approach has shown better tolerance to posture, occlusion, illumination or deformation. On the other hand, current methods usually involve spatio-temporal analysis of 3D descriptors at multiple scales in high resolution videos, so they require high computational costs.

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 332–339, 2011.

Approaches based on feature descriptors typically use the well-known bag-of-words framework [14,7]. In this case the output is simply a histogram that reflects *word* distribution as frequencies. In order to obtain the histogram, the bag-of-words representation creates a vocabulary using for example k-means clustering. The complete procedure is described in Section 4.

The goal of this paper is to assess the performance of modern action recognition approaches for the recognition of fights in videos, movies or video-surveillance footage. Most of previous work on action recognition focuses on simple human actions like walking, jumping or hand waving [13]. Despite its potential usefulness, violent action detection has been less studied. Whereas there is a number of well-studied datasets for action recognition, significant datasets with violent actions have not been made available. In this work we introduce a fight dataset and use two of the best action recognition methods currently available (STIP [12] and MoSIFT [4]) to assess the performance in the fight detection problem.

A violence detector has immediate applicability both in the surveillance domain and for rating/tagging online video content. The primary function of large-scale surveillance systems deployed in institutions such as schools, prisons and elder 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. The primary contribution of this paper are two-fold. First, we show that one can construct a versatile and accurate fight detector using a local descriptors approach. Second, we present a new dataset of hockey video containing fights and demonstrate that our proposed approach can reliably detect violence in sports footage, even in the presence of camera motion.

The paper is organized as follows. Section 2 analyzes previous work on violence recognition. Next, we describe the new hockey fights dataset in Section 3. Section 4 presents the two descriptors we use for activity recognition. Then, we describe the bag-of-words approach and the discriminative classifier. Section 5 details our evaluation methodology and summarizes our experimental results on the hockey fights dataset. Finally, in Section 6 we summarize key conclusions.

2 Related Work

One of the first proposals for violence recognition in video is Nam *et al.* [18], which proposes 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.* [5] 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.* [10] also propose a violence detector based on audio features. Clarin *et al.* [6] 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.* [19], introduce the CASSANDRA system, which employs motion features related to articulation in video and scream-like cues in audio to detect aggression in surveillance videos.

More recently, Gong *et al.* [11] 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.* [3] use binary local motion descriptors (spatio-temporal video cubes) and a bag-of-words approach to detect aggressive behaviors. Lin and Wang [15] 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.* [9] 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.

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 is not available and where the video is greyscale. Finally, while explosions, blood and running may be useful cues for violence in action movies, they are rare in real-world surveillance video. In this paper, we focus on reliable cues for early detection of violence in such settings.

3 Dataset

The majority of widely used, publicly-available datasets in action recognition, such as KTH [13], focus on single actors performing a simple action like walking, jumping or waving against an uncluttered background; these are clearly unsuitable for evaluating violence detection. Datasets such as INRIA IXMAS, which show an individual kicking or punching could be used to train (but not evaluate) fight detection systems. Some datasets like CAVIAR, BEHAVE or CareMedia contain some instances of people engaged in aggressive behaviors, but that is not their primary focus.

Our intention is to introduce a new video dataset created specifically for evaluating violence detection systems, where both normal and violent activities occur in similar, dynamic settings. To this end, we collected 1000 clips of action from hockey games of the National Hockey League (NHL), as shown in Fig. 1. Each clip consists of 50 frames of 720×576 pixels and is manually labeled as "fight" or "non-fight". This dataset enables us to easily and robustly measure the performance of a variety of violence recognition approaches, as shown in Section 5. Our fight dataset is available by request from the authors.

4 Activity Recognition

Local image features or interest points provide compact and abstract representations of patterns in an image. Analogously, with local spatio-temporal features

Fig. 1. Sample of a fight clip from our 1000-video database

it is possible to obtain compact and descriptive representations of motion. In this respect, two prominent spatio-temporal descriptors are Space-Time Interest Points (STIP) and Motion SIFT (MoSIFT).

As described in [12], Space-Time Interest Points (STIP) is an extension of the Harris corner detection operator to space-time. The detected interest points are characterized by a high variation of the intensity in space, and non-constant motion in time. These salient points are detected at multiple spatial and temporal scales. Then, HOG (Histograms of Oriented Gradients), HOF (Histograms of Optical Flow) and a combination of HOG and HOF termed HNF feature vectors are extracted for 3D video patches in the neighborhood of the detected STIPs. These features can be used for recognizing motion events with high performance and they are robust to scale, frequency and velocity variations of the pattern.

MoSIFT [4] is an extension of the popular SIFT [17] image descriptor for video. The standard SIFT extracts histograms of oriented gradients in the image. The 256-dimensional MoSIFT descriptor consists of two portions: a standard SIFT image descriptor and an analogous histogram of optical flows, which represents local motion. These descriptors are extracted only from regions of the image with sufficient motion. The MoSIFT descriptor has shown better performance in recognition accuracy than other state-of-the-art descriptors [4] but the approach is significantly more computationally expensive than STIP.

On the other hand, the Bag-of-Words (BoW) approach, adopted from the text retrieval community [14], has recently become popular for image [7] and video understanding [16]. The approach represents each video sequence as a histogram over a set of *visual words* to generate a fixed-dimensional encoding that can be processed using a standard classifier. In a learning phase, the vocabulary of visual words is typically defined as the cluster centers obtained from k-means clustering over a large collection of sample low-level descriptors (STIP or MoSIFT descriptors, see above). In our study, we evaluated vocabularies with 50, 100, 150, 200, 300, 500 and 1000 cluster centers.

Given a vocabulary, the next step is to quantize each descriptor extracted from the given video to the closest *visual word*, thus generating histograms of word occurrence. The final step of this BoW approach is the classification of the histograms. These histograms are high-dimensional vectors that can be classified using a standard classifier, typically a Support Vector Machine (SVM). It is well-known that the choice of SVM kernel can significantly affect performance; in our experiments, we explore the following popular kernels: the histogram intersection kernel [1] (HIK), radial basis function (RBF) and Chi-Squared kernel, which is a variant of RBF that uses χ^2 distance. Kernel parameters are tuned using 5-fold cross-validation.

5 Experimental Results

The BoW approach using the STIP and MoSIFT descriptors was evaluated on the 1000-clip hockey fight dataset. In order to assess the impact of vocabulary size, we generated vocabularies of 50, 100, 150, 200, 300, 500 and 1000 words. Table 1 presents the accuracy of fight detection, averaged over 5-fold cross-validation. For space reasons, we only show here results obtained with the histogram intersection kernel (HIK) since it consistently outperformed RBF and χ^2. We see that the BoW variants all achieve accuracies near 90%, with a slight improvement with increasing vocabulary size. On this dataset, STIP(HOG) and MoSIFT perform comparably. The ROC curve for the best of those runs is shown in Figure 2. Note that this result (MoSIFT on 500-word vocabulary) does not correspond to the highest result in Table 1 since the latter shows accuracies averaged over all folds.

Table 1. Accuracy of fight detection on 1000-clip hockey dataset (5-fold CV)

Vocabulary	STIP (HOG) + HIK	STIP (HOF) + HIK	MoSIFT + HIK
50	87.8%	83.5%	87.5%
100	89.1%	84.3%	89.4%
150	89.7%	85.9%	89.5%
200	89.4%	87.5%	90.4%
300	90.8%	87.2%	90.4%
500	91.4%	87.4%	90.5%
1000	**91.7**%	**88.6**%	**90.9**%

Hockey fights contain useful information for learning fight patterns. Still, can those patterns be translated to other scenarios? To explore the generalization capacity of the studied approaches, we also evaluated the fight recognition system on a second dataset consisting of 200 video clips obtained from action movies (see Figure 3 for examples), of which 100 contained a fight. 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. Table 2

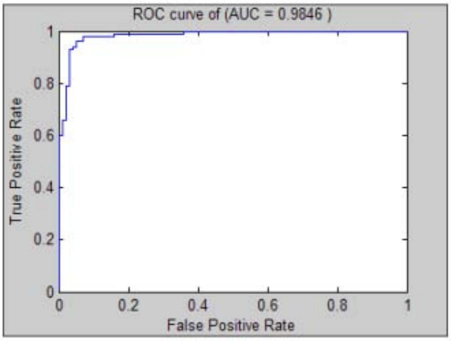

Fig. 2. ROC curve of fight detection on 1000-video Hockey dataset using MoSIFT with 500-word vocabulary and histogram intersection kernel

summarizes the results. In this case, STIP (HOF) outperformed STIP (HOG), but STIP's overall performance is very poor as compared to MoSIFT (59.0% vs. 89.5%). Note that increasing vocabulary size does not uniformly improve recognition accuracy.

We make several observations based on these experiments. First, detecting fights in televised hockey footage is easier than detecting fights in action movies, despite the fact that the former contains very similar footage for both classes. This could partially be attributed to the fact that fights in movies are more varied in appearance and cinematography while sports footage is relatively consistent. However, it also indicates that televised hockey fights may exhibit more reliable cues that a supervised classifier can exploit — for instance, the camera tends to zoom in to a hockey fight while showing more wide-angle shots during non-fight segments of the hockey game. Second, we see that STIP and MoSIFT are similar in performance on the former task but MoSIFT is dramatically superior

Fig. 3. Example of a fight sequence from an action movie

Table 2. Accuracy of fight detection on action movie dataset (5-fold CV)

Vocabulary	STIP (HOG) + HIK	STIP (HOF) + HIK	MoSIFT + HIK
50	44.5%	51.2%	76.0%
100	45.0%	56.5%	79.5%
150	**49.0**%	**59.0**%	80.0%
200	46.5%	53.5%	80.0%
300	44.5%	52.5%	87.5%
500	44.5%	50.5%	**89.5**%
1000	38.5%	52.5%	89.0%

on the action movie dataset (retaining 90% accuracy levels). This indicates that the MoSIFT representation, though more computationaly expensive than STIP, does make a difference.

6 Conclusions

Recognizing fights and aggressive behavior in video is an increasingly important application area. Such capability may be extremely useful in video surveillance scenarios like in prisons, psychiatric or elderly centers. Action recognition techniques that have focused largely on individual actors and simple events can be extended to this specific application. This paper evaluates how state-of-the-art video descriptors can perform fight detection on two new datasets: a 1000-video collection of NHL hockey games and a smaller 200-clip collection of scenes from action movies. Experiments show that the popular bag-of-words approach can accurately recognize fight sequences with approximately 90% accuracy. For the hockey dataset, we observed that accuracy was insensitive to the choice of low-level feature descriptor and vocabulary size; however, on the second dataset, the choice of descriptor was critical, with MoSIFT dramatically outperforming the best STIP under all conditions. The promising performance of action recognition methods on this challenging task shows that a versatile marketable fight detector may be feasible.

Acknowledgements. This work was partially funded by Project PII2I09-0043-3364 of Castilla-La Mancha Regional Government. We thank Drs. Alex Hauptmann and Ming-yu Chen for providing the MoSIFT descriptor and their helpful suggestions.

References

1. Barla, A., Odone, F., Verri, A.: Histogram intersection kernel for image classification. In: Proceedings of ICIP, pp. 513–516 (2003)
2. Bregler, C.: Learning and recognizing human dynamics in video sequences. In: Proceedings of Computer Vision and Pattern Recognition (1997)

3. Chen, D., Wactlar, H., Chen, M., Gao, C., Bharucha, A., Hauptmann, A.: Recognition of aggressive human behavior using binary local motion descriptors. In: Engineering in Medicine and Biology Society, pp. 5238–5241 (20-25 2008)
4. Chen, M., Hauptmann, A.: MoSIFT: Recognizing human actions in surveillance videos. Tech. rep., Carnegie Mellon University, Pittsburgh, USA (2009)
5. Cheng, W.H., Chu, W.T., Wu, J.L.: Semantic context detection based on hierarchical audio models. In: Proceedings of the ACM SIGMM workshop on Multimedia information retrieval, pp. 109–115 (2003)
6. Clarin, C., Dionisio, J., Echavez, M., Naval, P.C.: DOVE: Detection of movie violence using motion intensity analysis on skin and blood. Tech. rep., University of the Philippines (2005)
7. Csurka, G., Dance, C., Fan, L.X., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Workshop on Statistical Learning in Computer Vision (2004)
8. Efros, A.A., Berg, A.C., Mori, G., Malik, J.: Recognizing action at a distance. In: IEEE International Conference on Computer Vision, pp. 726–733 (2003)
9. Giannakopoulos, T., Makris, A., Kosmopoulos, D., Perantonis, S., Theodoridis, S.: Audio-visual fusion for detecting violent scenes in videos. In: Konstantopoulos, S., Perantonis, S., Karkaletsis, V., Spyropoulos, C.D., Vouros, G. (eds.) SETN 2010. LNCS, vol. 6040, pp. 91–100. Springer, Heidelberg (2010)
10. Giannakopoulos, T., Kosmopoulos, D., Aristidou, A., Theodoridis, S.: Violence content classification using audio features. In: Antoniou, G., Potamias, G., Spyropoulos, C., Plexousakis, D. (eds.) SETN 2006. LNCS (LNAI), vol. 3955, pp. 502–507. Springer, Heidelberg (2006)
11. Gong, Y., Wang, W., Jiang, S., Huang, Q., Gao, W.: Detecting violent scenes in movies by auditory and visual cues. In: Proceedings of the 9th Pacific Rim Conference on Multimedia, pp. 317–326. Springer, Heidelberg (2008)
12. Laptev, I.: On space-time interest points. International Journal of Computer Vision 64, 107–123 (2005)
13. Laptev, I., Lindeberg, T.: Space-time interest points. In: Proceedings of International Conference on Computer Vision, pp. 432–439 (2003)
14. Lewis, D.: Naive Bayes at Forty: The independence assumption in information retrieval. In: European Conference on Machine Learning, pp. 4–15 (1998)
15. Lin, J., Wang, W.: Weakly-supervised violence detection in movies with audio and video based co-training. In: Muneesawang, P., Wu, F., Kumazawa, I., Roeksabutr, A., Liao, M., Tang, X. (eds.) PCM 2009. LNCS, vol. 5879, pp. 930–935. Springer, Heidelberg (2009)
16. Lopes, A.P.B., do Valle Jr., E.A., de Almeida, J.M., de Albuquerque Araújo, A.: Action recognition in videos: from motion capture labs to the web. CoRR abs/1006.3506 (2010)
17. Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(91) (2004)
18. Nam, J., Alghoniemy, M., Tewfik, A.: Audio-visual content-based violent scene characterization. In: Proceedings of ICIP, pp. 353–357 (1998)
19. Zajdel, W., Krijnders, J., Andringa, T., Gavrila, D.: CASSANDRA: audio-video sensor fusion for aggression detection. In: IEEE Conference on Advanced Video and Signal Based Surveillance, AVSS 2007, pp. 200–205 (2007)