Fight Recognition in Video Using Hough Forests and 2D Convolutional Neural Network

Ismael Serrano, Oscar Deniz, *Senior Member, IEEE*, Jose Luis Espinosa-Aranda, and Gloria Bueno, *Member, IEEE*

***Abstract*— While action recognition has become an important line of research in computer vision, the recognition of particular events, such as aggressive behaviors, or fights, has been relatively less studied. These tasks may be extremely useful in several video surveillance scenarios, such as psychiatric wards, prisons, or even in personal camera smartphones. Their potential usability has led to a surge of interest in developing fight or violence detectors. One of the key aspects in this case is efficiency, that is, these methods should be computationally fast. "Handcrafted" spatio-temporal features that account for both motion and appearance information can achieve high accuracy rates, albeit the computational cost of extracting some of those features is still prohibitive for practical applications. The deep learning paradigm has been recently applied for the first time to this task too, in the form of a 3D convolutional neural network that processes the whole video sequence as input. However, results in human perception of other's actions suggest that, in this specific task, motion features are crucial. This means that using the whole video as input may add both redundancy and noise in the learning process. In this paper, we propose a hybrid "handcrafted/learned" feature framework which provides better accuracy than the previous feature learning method, with similar computational efficiency. The proposed method is compared to three related benchmark data sets. The method outperforms the different state-of-the-art methods in two of the three considered benchmark data sets.**

***Index Terms*— Fight recognition, violence recognition, Hough forests, deep learning, 2D convolutional neuronal network.**

I. Introduction

In RECENT years, the task of human action recognition from video has been tackled with computer vision and machine learning techniques, see surveys [1]–[3]. Experimental results have been obtained for recognition of actions such as walking, jogging, pointing or hand waving [4]. However, action detection has been devoted comparatively less effort. Violence detection is a task that can be leveraged in real-life applications. While there is a large number of studied datasets for action recognition, specific datasets with a relevant number of violent sequences (fights) were not available until [5], where the authors created two specific datasets for the fight/violence problem testing state-of-the-art methods on them. The main task of large-scale surveillance systems used in institutions such as prisons, schools and psychiatric care facilities is generating alarms of potentially dangerous situations. Nevertheless, security guards are frequently burdened with a large number of cameras where manual response times are frequently large, resulting in a strong demand for automated alert systems. Also, this type of systems must be very efficient because there is generally a large number of surveillance cameras which must be processed. Similarly, there is increasing demand for automated rating and tagging systems that can process large amounts of videos uploaded to websites. Since smartphones are often used to record beatings, efficient mobile implementations are desired too.

This work is based on the assumption that fights in video can be reliably recognized by kinematic cues that represent violent motion. This idea is inspired by a body of research on human perception that has shown that the kinematic pattern of movement is sufficient for the perception of other's actions [6]. 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 [7]–[10], albeit the degree of correlation varies for different emotions. Thus, features such as acceleration and jerkiness tend to be associated with emotions with high activation (e.g. anger, happiness), whereas slow and smooth movements are more likely to be judged as emotions with low activation (eg. sadness). This same essential idea has been also supported by research on the computer vision side [11], [12]. These authors demonstrate that kinematic patterns of movements and dynamic features are representative for the perception of high-energy actions. In other words, motion carries most of the information useful to discriminate fight/violence sequences. Moreover, motion information could be much more important than appearance in this task. Following these experiments we propose to leverage the high-motion areas in this type of sequences using spatial features combined with a spatio-temporal classifier to learn when and where the fight/violence actions could be occurring.

Still, in line with the classical approach to machine learning, "handcrafted" features have been mostly used in previous work related to this task. In this work, features are learned using a Convolutional Neural Network trained with images that summarize the content of video sequences. In this respect, the proposed method can be considered hybrid in the sense that

Manuscript received June 28, 2017; revised February 10, 2018 and March 26, 2018; accepted June 5, 2018. Date of publication June 8, 2018; date of current version June 27, 2018. This work was supported by the Spain's Ministry of Economy and Competitiveness under Project TIN2011-24367. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Alin M. Achim. *(Corresponding author: Ismael Serrano.)*

The authors are with the VISILAB Group, E. T. S. I. Industriales, University of Castilla–La Mancha, 13071 Ciudad Real, Spain (e-mail: Ismael.Serrano@uclm.es).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2018.2845742

[39] J. W. Davis and A. F. Bobick, "The representation and recognition of human movement using temporal templates," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 1997, pp. 928–934.
[40] S. Leutenegger, M. Chli, and R. Y. Siegwart, "BRISK: Binary robust invariant scalable keypoints," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 2548–2555.
[41] E. Rosten and T. Drummond, "Fusing points and lines for high performance tracking," in *Proc. 10th IEEE Int. Conf. Comput. Vis.*, vol. 2, Oct. 2005, pp. 1508–1515.
[42] J. Gall, A. Yao, N. Razavi, L. Van Gool, and V. Lempitsky, "Hough forests for object detection, tracking, and action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 11, pp. 2188–2202, Nov. 2011.
[43] A. Yao, J. Gall, and L. Van Gool, "A Hough transform-based voting framework for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 2061–2068.
[44] G. Garcia-Hernando, H. J. Chang, I. Serrano, O. Deniz, and T.-K. Kim, "Transition Hough forest for trajectory-based action recognition," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–8.
[45] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[46] E. Parzen, "On estimation of a probability density function and mode," *Ann. Math. Statist.*, vol. 33, no. 3, pp. 1065–1076, Sep. 1962.
[47] S. Blunsden and B. Fisher, "The BEHAVE video dataset: Ground truthed video for multi-person behavior classification," *Ann. BMVA*, vol. 4, no. 4, pp. 1–11, 2010.
[48] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. Int. Conf. Artif. Intell. Statist. Soc.*, 2010, pp. 249–256.
[49] T. Zhang, Z. Yang, W. Jia, B. Yang, J. Yang, and X. He, "A new method for violence detection in surveillance scenes," *Multimedia Tools Appl.*, vol. 75, no. 12, pp. 7327–7349, 2016.
[50] D. Xu, E. Ricci, Y. Yan, J. Song, and N. Sebe. (2015). "Learning deep representations of appearance and motion for anomalous event detection." [Online]. Available: https://arxiv.org/abs/1510.01553
[51] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.
[52] T. Zhang, W. Jia, B. Yang, J. Yang, X. He, and Z. Zheng, "MoWLD: A robust motion image descriptor for violence detection," *Multimedia Tools Appl.*, vol. 76, no. 1, pp. 1419–1438, 2017.
[53] I. Serrano, O. Deniz, G. Bueno, G. Garcia-Hernando, and T.-K. Kim, "Spatio-temporal elastic cuboid trajectories for efficient fight recognition using Hough forests," *Mach. Vis. Appl.*, vol. 29, no. 2, pp. 207–217, 2018.
[54] H. Wang, A. Kläser, C. Schmid, and C.-L. Lin, "Action recognition by dense trajectories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Colorado Springs, CO, USA, Jun. 2011, pp. 3169–3176.

Ismael Serrano received the degree in computer science and the Ph.D. degree *(cum laude)* from the University of Castilla–La Mancha, Spain, in 2012 and 2016, respectively. His Ph.D. thesis was on fight detection in video using computer vision and machine learning techniques. He scored the highest mark in his final degree project about person detection. He has served as a Researcher Collaborator with Imperial College London, U.K., and Leica Biosystems, Ireland. He is currently a Ph.D. Researcher with the Industry and Advanced Manufacturing Department, Vicomtech Foundation. He has authored over 15 papers in journals and conferences. He has published two books on OpenCV. His research interests are mainly focused on computer vision and machine learning, especially on deep learning.

Oscar Deniz (SM'98) has been a Visiting Researcher with Carnegie Mellon University, USA, Imperial College London, U.K., and Leica Biosystems, Ireland. He is currently an Associate Professor with the University of Castilla La–Mancha and contributes to VISILAB. He is the Coordinator of the European H2020 Project Eyes of Things and a Partner in the European H2020 Project BONSEYES. He has authored over 50 refereed papers in journals and conferences. His research interests are mainly focused on computer vision and pattern recognition. He is with the AAAI, SIANI, CEA-IFAC, AEPIA, AERFAI-IAPR, and The Computer Vision Foundation. He serves as an Academic Editor of *PLOS One* journal. He is a Reviewer/Technical Expert for EU programs and an Advisory Board Member of the H2020 Project TULIPP.

Jose Luis Espinosa-Aranda received the degree in computer engineering and the Ph.D. degree in computer science from the University of Castilla–La Mancha, Spain, in 2009 and 2014, respectively. He is currently a Research Assistant of the VISILAB Group and also an Associate Professor with the University of Castilla–La Mancha. His current research interests include artificial intelligence and computer vision.

Gloria Bueno (M'99) received the degree in physics from UCM, Madrid, Spain, in 1993, and the Ph.D. degree in machine vision from Coventry University, U.K., in 1998. She has served as a Visiting Researcher with Carnegie Mellon University, USA, and Leica Biosystems, Ireland. She has experience as a principal researcher in several research centers, such as CNRS, Louis Pasteur University, Strasbourg, France, Gilbert Gilkes & Gordon Technology, U.K., and CEIT San Sebastian, Spain. She is currently a Professor with the Engineering School, University of Castilla–La Mancha. She has been a principal researcher of different national and international projects focused on artificial intelligence and image processing. She is the Coordinator of the European AIDPATH Project—Academia and Industry Collaboration for Digital Pathology. She has authored two patents, four registered software, and over 80 refereed papers. She is with several societies, such as ESDIP, CEA-IFAC, SEMF, and SEIB.