Handcrafted vs. Learned Features for Automatically Detecting Violence in Surveillance Footage

Arnaldo V. Barros da Silva[1], **Luis F. Alves Pereira**[1]

[1]Universidade Federal do Agreste de Pernambuco (UFAPE)
55292-270 – Garanhuns – PE – Brazil

{arnaldovitorbarros@gmail.com, luis-filipe.pereira.ufape.edu.br}

Abstract. *For many years, methods for detecting violence in video data used features designed by humans to extract visual information from input frames for composing feature vectors and then applied machine learning techniques to assign labels to them. Recently, Deep Learning methods are highly evidenced for this task since they can automatically learn image features. Furthermore, they usually overcome the accuracy rates obtained by classical methods based on handcrafted features. This work evaluates learned and handcrafted features for classifying video frames as 'violence' or 'non-violence'. Our results showed that learned features can not always be claimed superior since some violent scenes are only detected by handcrafted features.*

1. Introduction

As the number of surveillance cameras installed worldwide increases, the urgency for real-time video-content analysis also grows. Once supervising multiple monitors for a long time is an unsuited task for human agents, many computer vision algorithms have been proposed during the last decade for detecting abnormal and potentially dangerous situations, such as: *(i)* people disobedience to virtual fence [Delgado et al. 2014, Chen et al. 2012]; *(ii)* people loitering [Coşar et al. 2016, Arroyo et al. 2015]; *(iii)* crowd panic [Krausz and Bauckhage 2012, Zhang et al. 2019]; *(iv)* seniors falling [Lu et al. 2018, Rougier et al. 2011]; and others. Detecting violent scenes is also relevant for ensuring safety in public areas and private properties, thus guaranteeing stability in people's lives and gradually allowing a safer society.

Classical methods for detecting violence in video data used image handcrafted features such as optical flow [Gao et al. 2016], appearance [Chen and Hauptmann 2009], and acceleration [Deniz et al. 2014] to compose feature vectors. Then, traditional machine learning techniques such as Support Vector Machine (SVM) [Hearst et al. 1998] were used for assigning each of those feature vectors to the labels 'violence' and 'non-violence'. Using such an approach, it was possible, for instance, to detect violent scenes from videos of hockey matches with accuracy rates above 85% [Laptev et al. 2008].

By the end of the 2000s, Deep Learning techniques emerged within a new branch of machine learning [Yang et al. 2015]. They eliminated the need for humans to design features as they learned to extract image features in the first layers of Convolutional Neural Networks (CNN's). Thus, CNN's has become increasingly popular, reporting near 100% accuracy rates for detecting violent scenes from movies and hockey matches using complex image characteristics to human interpretation [Zhou et al. 2017, Soliman et al. 2019, Keçeli and Kaya 2017].

Despite the massive success of Deep Learning methods for violence detection, one question related to their performance is still not entirely explained: *are both learned and handcrafted features focusing on the same aspects of the images?* For this reason, we designed a framework not only for evaluating the accuracy of handcrafted and learned features but also for exploring the data visualization in order to study the datasets separability. Moreover, we executed experiments to evaluate whether the sets of correctly classified videos by both techniques differ or not. A divergence between those sets indicates that the methods would focus differently within the video frames.

This paper is organized as follows: Section 2 describes the related works; Section 3 describes the framework designed to evaluate the handcrafted ans learned features; Section 4 describes dataset and the experimental parameters; Section 5 presents the results obtained and discusses them; finally, Section 6 concludes the paper.

2. Related Works

Among the classical methods based on handcrafted features, we can mention: i) the STIP [De Souza et al. 2010], which considered local spatio-temporal features within bags of visual words to construct feature vectors; ii) the RIMOC [Ribeiro et al. 2016], which was created from the eigenvalues obtained from the Optical Flow Histogram (HOF) extracted in consecutive instants of time embedded in a spherical Riemannian manifold; iii) the MoSIFT [Chen and Hauptmann 2009], which had feature vectors generated by concatenating the Oriented Gradient Histogram (HOG) to the HOF. The MoSIFT method will be explained in more depth in the following sections as it was chosen to compose the handcrafted feature extractors of our framework.

With respect to the methods based on Deep Learning that eliminated the requirement for expert-based handcrafted features to perform violence detection, we highlight i) the multi-stream deep neural network where raw videos, optical streams, and acceleration stream maps are given as inputs to three network branches, then there is a fusion of those branches using a Long Short Term Memory (LSTM) as proposed by Dong et al. [Dong et al. 2016]; ii) the usage of a CNN to extract frame-level features from the video frames, and posterior feature aggregation using a variant of LSTM-based network as proposed by Sudhakaran et al. [Sudhakaran and Lanz 2017]; iii) the Flow Gated Network, where a 3D CNN is trained using multiple video frames in sequence, and another 3D CNN is trained using the optical flow of the respective frames, then the two branches are combined using temporal pooling as proposed by Cheng et al. [Cheng et al. 2020].

Our work is also related to other papers that conducted similar evaluation between handcrafted and learned features for different image classification problems, as follows: i) Saba [Saba 2021] makes a similar comparison in the context of skin cancer detection. He found cases where the handcrafted features reached accuracy rates higher than the learned features, reaching up to 100% accuracy in one of the evaluated datasets; ii) Already Antipov et al. [Antipov et al. 2015] evaluated the use of both types of features for the problem of gender recognition of pedestrians. Their study showed that both approaches have similar performance in small homogenous data, but the handcrafted features underperformed the learned features in mean average precision for more complex datasets; iii) Nanni et al. [Nanni et al. 2017] did an extensive comparison between handcrafted and learned extractors using image datasets related to many image

classification problems. Their experiments have shown that there are indeed contexts where handcrafted can overcome the learned features concerning accuracy. None of those works, however, investigated whether there is a divergence between the image aspects evaluated by both handcrafted and learned features.

3. Methods

The framework applied to evaluate different subsets of features is illustrated in Figure 1. It is mainly based on the bag-of-features technique [Nowak et al. 2006] associated with two types of image descriptors: *(i)* traditional handcrafted features and *(ii)* learned features automatically extracted using Deep Learning models, such as the VGG-19 [Simonyan and Zisserman 2014]. Using the bag-of-features technique, each video containing any number of frames is easily associated with a single feature vector in the space $\mathbf{U}$. Thus, by applying a technique for dimensionality reduction, *e.g.* TSNE [Maaten and Hinton 2008], it is possible to *visualize* the dataset and have an intuition about its classification challenge.

For a given dataset with M videos, let: *(i)* N_m be the number of frames in the m^{th} video, *(ii)* $T1$, $T2$, $T3$ be three sets of integers that refer to indexes of videos in a dictionary set, a training set and a test set, respectively. Then, the frames $\{\mathbf{x}_{mn}\}, \forall mp \in T1, \forall n \in \{0, 1, ..., N_m\}$ are processed by a feature extractor to compose a space of features F. Next, by clustering such space within w words, a dictionary $\mathbf{D}$ is created. After that, for each $m \in \{T2, T3\}$, the frames $\{\mathbf{x}_{mn}\}, \forall n \in \{0, 1, ..., N_m\}$ are processed by the feature extractor and the dictionary D [Jégou et al. 2009] to create new vectors in the space of videos $\mathbf{U}$. Finally, the performance of a classifier C is computed over the data in the space $\mathbf{U}$.

With respect to the *feature extraction* modules presented in the proposed framework, we implemented them using two approaches based on handcrafted and learned features described in the following subsections.

3.1. The handcrafted feature extractor

It uses the MoSIFT [Chen and Hauptmann 2009] technique to obtain a set of feature vectors for each input frame. This processing is described by the Algorithm 1. First, SIFT keypoints [Lowe 2004] are computed for each frame input to find regions of interest. Then, a MoSIFT vector of size 256 is created by concatenating SIFT and HOF [Van Gool 2008] descriptors for those regions of interest which have optical flow [Farnebäck 2003] greater than a threshold ϵ.

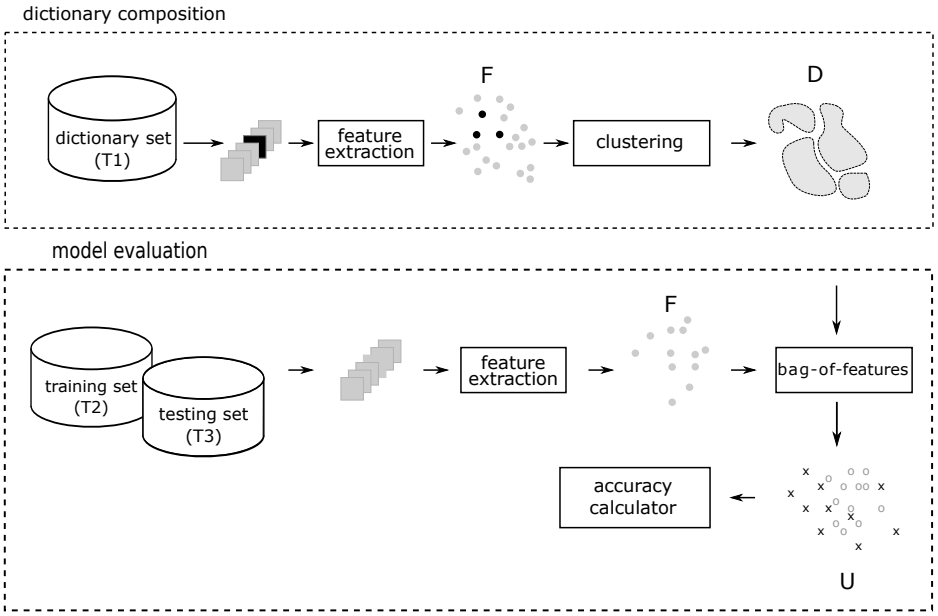

Figure 1. The evaluation framework applied in our study uses the dictionary set $T1$, the training set $T2$, and the test set $T3$. The feature extractor processes videos from $T1$ to create a feature space F, where multiple vectors refer to a single video frame. Next, by clustering F in w words, a dictionary D is created. Then, using D and the bag-of-features technique, the data from $T2$ and $T3$ are converted into a video space U. In U, each vector refers to a single video clip. Finally, the classification accuracy of a classifier C is measured over the videos in U.

Algorithm 1: The handcrafted feature extractor

Input: $frame, next\_frame, \epsilon$
Output: $hand\_features$
$hand\_features \leftarrow []$
$keypoints \leftarrow SIFT(frame)$
for *each* $kp \in keypoints$ **do**
  if $opticalFlow(frame, next\_frame, kp.position) > \epsilon$ **then**
    $hof \leftarrow HOF(frame, next\_frame, kp.position)$
    $mosift \leftarrow cat(hof, kp.descriptor)$
    $hand\_features.add(mosift)$
  end
end

3.2. The learned feature extractor

It uses a VGG-19 [Simonyan and Zisserman 2014] (illustrated in Figure 2) to extract features from the video frames and its optical flow, as suggested by Xu *et al.* [Xu et al. 2017]. A detailed description of the entire feature extractor is presented in Algorithm 2. A pre-trained instance of the VGG-19 is used to obtain two feature vectors of size $4,096$ each: one related to the RGB frame and the other related to its optical flow.

Algorithm 2: The learned feature extractor

Input: $frame, next\_frame$

Output: $learned\_features$

$learned\_features \leftarrow []$

$flow \leftarrow optical\_flow(frame, next\_frame)$

$learned\_features.add(vgg\_19(frame))$

$learned\_features.add(vgg\_19(flow))$

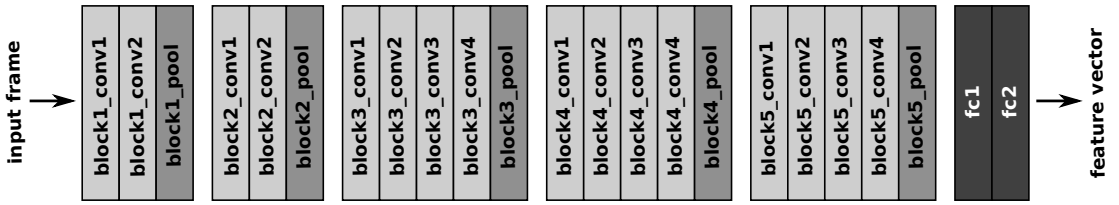

Figure 2. The architecture of the VGG-19 used in this work to compose the *learned feature extractor*. A 224×224 input image propagates through 16 convolutional and 2 fully connected layers to generate a feature vector of size $4,096$.

4. Experiments

4.1. Datasets

In our experiments, we use three datasets: i) the Hockey Fight dataset [Nievas et al. 2011] which contains 1000 clips captured from matches of the National Hockey League (NHL) manually labeled as *fight* or *non-fight*; ii) the Violent Flows dataset [Hassner et al. 2012] which is composed of violent and non-violent crowd behavior in real-world footage collected at YouTube and LiveLeak, this benchmark comprises 246 clips that include specters at large events, people protesting in the streets, and others and iii) the RWF-2000 dataset [Cheng et al. 2020] which contains 2,000 videos captured by surveillance cameras in real-world scenes labeled as violence and non-violence. A comparison among the datasets can be seen in Table 1.

Table 1. Characteristics of the used datasets used in the experiments. The latest two lines present the features related to the datasets created in this work.

dataset	# violent	# non-violent	hours length	resource	release year
Hockey [Nievas et al. 2011]	500	500	0.44	sports	2011
Violent Flows [Hassner et al. 2012]	123	123	0.8	real-world outdoor	2012
RWF-2000 [Cheng et al. 2020]	1000	1000	2.8	real-world indoor and outdoor	2020

4.2. Experimental parameters

We conducted experiments using three partitions comprising 50%, 25%, and 25% of each dataset for generating the dictionary (T1), training (T2), and testing (T3), respectively.

The clustering algorithm to generate the space U was *k-means algorithm* [Forgy 1965]. The number of words w was $2,048$, and the optical flow threshold ϵ in Algorithm 1 was 0.5. The VGG-19 in the Algorithm 2 was loaded with the weights learned from the ImageNet dataset [Deng et al. 2009] as the literature shows it may provide good generalization results for many image domains [Wen et al. 2019, Alhindi et al. 2018]. Furthermore, the classifier C was a Fully Connected Network composed of three dense layers activated by ReLu containing $1,024$, 512, and 128 neurons, and two more neurons composing the final layer activated by sigmoid.

5. Results and Discussions

5.1. Feature evaluation

Visualizations of the space of videos $\mathbf{U}$ - reduced to 2 dimensions via TSNE [Maaten and Hinton 2008] - are shown in Figure 3 along three columns, one for each dataset: the Hockey [Nievas et al. 2011], the Violent Flows [Hassner et al. 2012] and the RWF-2000 [Cheng et al. 2020] respectively. The first and second lines show the space of videos $\mathbf{U}$ generated using the handcrafted and the learned feature extractor. According to Figure 3, it is clear that the Hockey dataset [Nievas et al. 2011] has the highest inter-class and the lowest intra-class dispersion; as a consequence, it should be the easiest to classify. The Violent Flows [Hassner et al. 2012] also looks like a simple dataset, especially when extracted using learned features. The RWF-2000 [Cheng et al. 2020] contains footage of real-world violent crimes, and its space of videos $\mathbf{U}$, seems to be more complex as we would expect.

5.2. Accuracy

Table 2 shows the false-positive rate (FPR), the false-negative rate (FNR), and the accuracy rate (ACC) obtained by applying the proposed evaluation framework to all benchmarks evaluated in this work. As we would expect from the visualization of the video spaces U in Figure 3, the highest accuracy rates were obtained in the Hockey and Violent Flows datasets. Based on the literature, we would expect learned features to report superior classification rates than handcrafted features. This fact indeed happened for the Violent Flows and RWF-2000 datasets. However, for Hockey Fight - even by a slight difference - the handcrafted feature surpassed the learned feature.

5.3. Handcrafted versus Learned features

We cannot simply claim that the learned features are superior to the handcrafted features by observing only the accuracies. In order to conclude that, the set of videos correctly classified by the handcrafted features should be a subset of the set of videos correctly classified by the learned features. In contrast, the Venn diagrams Figure 4 show that the videos correctly classified by handcrafted and learned videos are different from each other. This result indicates that both types of features focus on different aspects of the video frames. From the Venn diagrams Figure 4, we can calculate the rate of misclassified videos by both handcrafted and learned features. This rate in the RWF-2000 dataset is 10%, the highest among all the evaluated benchmarks.

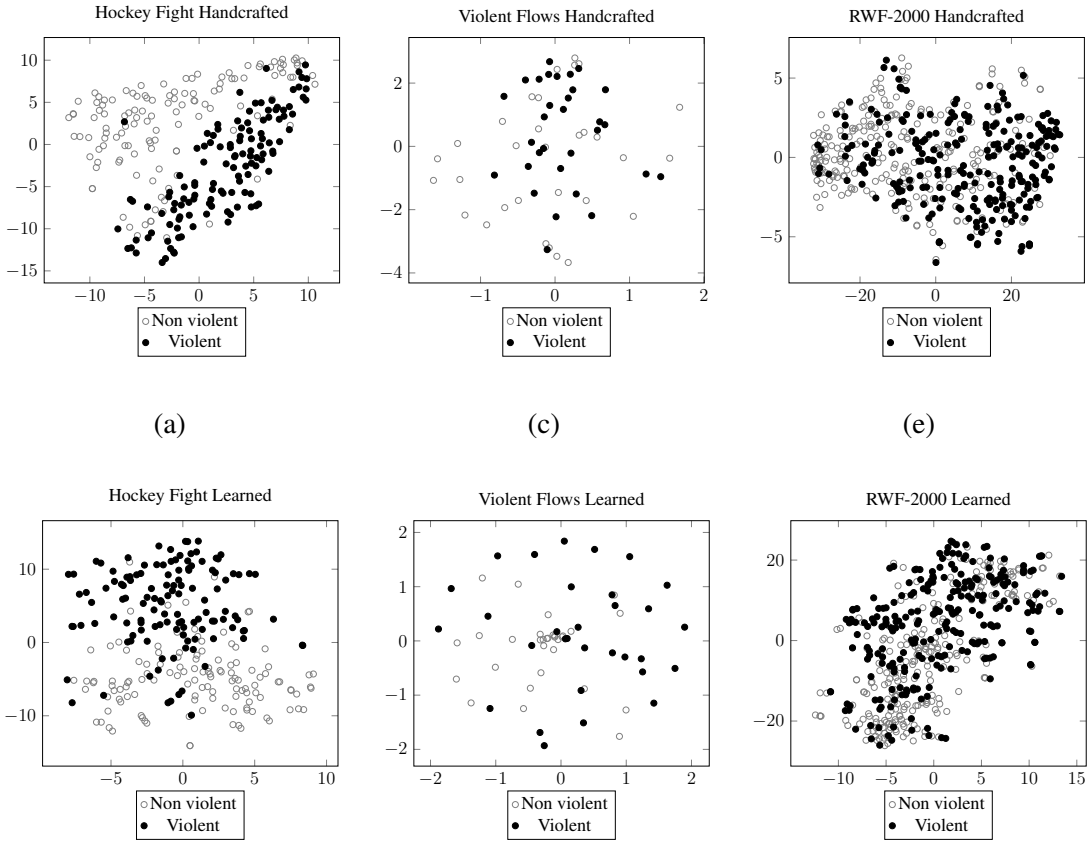

Figure 3. Spaces of videos U reduced to two dimensions via TSNE. Each column contains data of a different video collection extracted using handcrafted (first row) and learned features (second row). Those representations show that the classification tasks involving videos containing real-world violence (e) to (f) are more complex than the others.

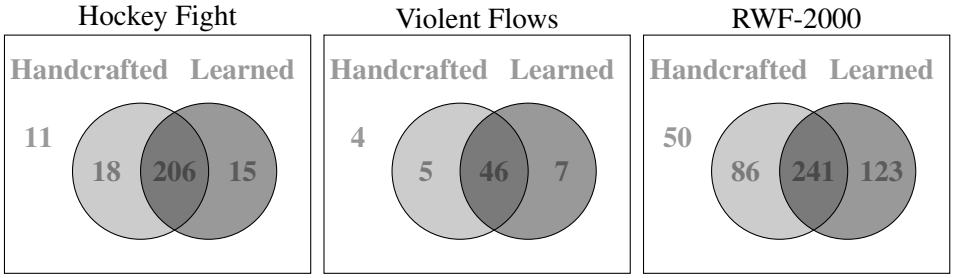

Figure 4. Number of samples corrected classified by the handcrafted and learned features.

6. Conclusion

Most modern solutions for violence detection use Deep Learning that avoids extracting handcrafted features from video frames. In this work, we compared solutions based on handcrafted and learned image features for detecting violent scenes in video frames. Our experiments indicated that the features learned by deep neural networks provide higher

Table 2. False-positive rate, false-negative rate, and accuracy rate obtained using the proposed evaluation framework in Hockey, Violent Flows and RWF-2000.

Dataset	Feature	FPR	FNR	ACC
Hockey [Nievas et al. 2011]	handcrafted	8.8%	12%	89.6%
Hockey [Nievas et al. 2011]	learned	7.2%	16%	88.4%
Violent Flows [Hassner et al. 2012]	handcrafted	12.9%	22.6%	82.2%
Violent Flows [Hassner et al. 2012]	learned	6.4%	22.5%	85.4%
RWF-2000	handcrafted	34.8%	34.4%	65.4%
RWF-2000	learned	21.2%	32.2%	72.8%

accuracy rates for detecting violence in videos. However, they do not seem to be able to fully replace the use of handcrafted features as we demonstrated that both types of features focus on different aspects of the input images and correctly classify different samples.

References

Alhindi, T. J., Kalra, S., Ng, K. H., Afrin, A., and Tizhoosh, H. R. (2018). Comparing lbp, hog and deep features for classification of histopathology images. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7.

Antipov, G., Berrani, S. A., Ruchaud, N., and Dugelay, J.-L. (2015). Learned vs. hand-crafted features for pedestrian gender recognition.

Arroyo, R., Yebes, J. J., Bergasa, L. M., Daza, I. G., and Almazán, J. (2015). Expert video-surveillance system for real-time detection of suspicious behaviors in shopping malls. *Expert systems with Applications*, 42(21):7991–8005.

Chen, J.-H., Tseng, T.-H., Lai, C.-L., and Hsieh, S.-T. (2012). An intelligent virtual fence security system for the detection of people invading. *9th International Conference on Ubiquitous Intelligence and Computing and 9th International Conference on Autonomic and Trusted Computing*, pages 786–791.

Chen, M.-y. and Hauptmann, A. (2009). Mosift: Recognizing human actions in surveillance videos.

Cheng, M., Cai, K., and Li, M. (2020). Rwf-2000: An open large scale video database for violence detection.

Coşar, S., Donatiello, G., Bogorny, V., Garate, C., Alvares, L. O., and Brémond, F. (2016). Toward abnormal trajectory and event detection in video surveillance. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(3):683–695.

De Souza, F. D., Chavez, G. C., do Valle Jr, E. A., and Araújo, A. d. A. (2010). Violence detection in video using spatio-temporal features. *23rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 224–230.

Delgado, B., Tahboub, K., and Delp, E. J. (2014). Automatic detection of abnormal human events on train platforms. *IEEE National Aerospace and Electronics Conference*, pages 169–173.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *IEEE conference on computer vision and pattern recognition*, pages 248–255.

Deniz, O., Serrano Gracia, I., Bueno, G., and Kim, T.-T. (2014). Fast violence detection in video. volume 2.

Dong, Z., Qin, J., and Wang, Y. (2016). Multi-stream deep networks for person to person violence detection in videos. *Chinese Conference on Pattern Recognition*, pages 517–531.

Farnebäck, G. (2003). Two-frame motion estimation based on polynomial expansion. *In: Image analysis*, 2749:363–370.

Forgy, E. W. (1965). Cluster analysis of multivariate data : efficiency versus interpretability of classifications. *Biometrics*, 21:768–769.

Gao, Y., Liu, H., Sun, X., Wang, C., and Liu, Y. (2016). Violence detection using oriented violent flows. *Image and vision computing*, 48:37–41.

Hassner, T., Itcher, Y., and Kliper-Gross, O. (2012). Violent flows: Real-time detection of violent crowd behavior. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1–6.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., and Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28.

Jégou, H., Douze, M., and Schmid, C. (2009). Packing bag-of-features. *IEEE 12th International Conference on Computer Vision*, pages 2357–2364.

Keçeli, A. and Kaya, A. (2017). Violent activity detection with transfer learning method. *Electronics Letters*, 53(15):1047–1048.

Krausz, B. and Bauckhage, C. (2012). Loveparade 2010: Automatic video analysis of a crowd disaster. *Computer Vision and Image Understanding*, 116(3):307–319.

Laptev, I., Marszalek, M., Schmid, C., and Rozenfeld, B. (2008). Learning realistic human actions from movies. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110.

Lu, N., Wu, Y., Feng, L., and Song, J. (2018). Deep learning for fall detection: Three-dimensional cnn combined with lstm on video kinematic data. *IEEE journal of biomedical and health informatics*, 23(1):314–323.

Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

Nanni, L., Ghidoni, S., and Brahnam, S. (2017). Handcrafted vs non-handcrafted features for computer vision classification. *Pattern Recognition*, 71.

Nievas, E. B., Suarez, O. D., García, G. B., and Sukthankar, R. (2011). Violence detection in video using computer vision techniques. *International conference on Computer analysis of images and patterns*, pages 332–339.

Nowak, E., Jurie, F., and Triggs, B. (2006). Sampling strategies for bag-of-features image classification. In *European conference on computer vision*, pages 490–503. Springer.

Ribeiro, P. C., Audigier, R., and Pham, Q. C. (2016). Rimoc, a feature to discriminate unstructured motions: Application to violence detection for video-surveillance. *Computer vision and image understanding*, 144:121–143.

Rougier, C., Meunier, J., St-Arnaud, A., and Rousseau, J. (2011). Robust video surveillance for fall detection based on human shape deformation. *IEEE Transactions on circuits and systems for video Technology*, 21(5):611–622.

Saba, T. (2021). Computer vision for microscopic skin cancer diagnosis using handcrafted and non-handcrafted features. *Microscopy Research and Technique*, 84:1272 – 1283.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Soliman, M. M., Kamal, M. H., El-Massih Nashed, M. A., Mostafa, Y. M., Chawky, B. S., and Khattab, D. (2019). Violence recognition from videos using deep learning techniques. In *2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS)*, pages 80–85.

Sudhakaran, S. and Lanz, O. (2017). Learning to detect violent videos using convolutional long short-term memory. *14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6.

Van Gool, L. (2008). Action snippets: How many frames does human action recognition require? *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*.

Wen, L., Li, X., Li, X., and Gao, L. (2019). A new transfer learning based on vgg-19 network for fault diagnosis. In *2019 IEEE 23rd International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 205–209.

Xu, D., Yan, Y., Ricci, E., and Sebe, N. (2017). Detecting anomalous events in videos by learning deep representations of appearance and motion. *Computer Vision and Image Understanding*, 156:117–127.

Yang, S., Luo, P., Loy, C. C., and Tang, X. (2015). From facial parts responses to face detection: A deep learning approach. pages 3676–3684.

Zhang, X., Shu, X., and He, Z. (2019). Crowd panic state detection using entropy of the distribution of enthalpy. *Physica A: Statistical Mechanics and its Applications*, 525:935–945.

Zhou, P., Ding, Q., Luo, H., and Hou, X. (2017). Violent interaction detection in video based on deep learning. *Journal of Physics: Conference Series*, 844:012044.