

Fight Detection from Still Images in the Wild

Şeymanur Akt

Istanbul Technical University

akti15@itu.edu.tr

Ferda O i

Qatar Computing Research Institute

fofli@hbku.edu.qa

Muhammad Imran

Qatar Computing Research Institute

mimran@hbku.edu.qa

Haz m Kemal Ekenel

Istanbul Technical University

ekenel@itu.edu.tr

Abstract

Detecting fights from still images shared on social media is an important task required to limit the distribution of violent scenes in order to prevent their negative effects. For this reason, in this study, we address the problem of fight detection from still images collected from the web and social media. We explore how well one can detect fights from just a single still image. We also propose a new dataset, named Social Media Fight Images (SMFI), comprising real-world images of fight actions. Results of the extensive experiments on the proposed dataset show that fight actions can be recognized successfully from still images. That is, even without exploiting the temporal information, it is possible to detect fights with high accuracy by utilizing appearance only. We also perform cross-dataset experiments to evaluate the representation capacity of the collected dataset. These experiments indicate that, as in the other computer vision problems, there exists a dataset bias for the fight recognition problem. Although the methods achieve close to 100% accuracy when trained and tested on the same fight dataset, the cross-dataset accuracies are significantly lower, i.e., around 70% when more representative datasets are used for training. SMFI dataset is found to be one of the two most representative datasets among the utilized fight datasets.

1. Introduction

With the rapid increase in social media usage, it is inevitable to face off the negative effects of certain types of content shared on these platforms, including but not limited to violent scenes, crime scenes, fight scenes, and scenes with dismembered body parts, among others. Such content, in the form of images and videos, especially for the younger users, can be inconvenient and harmful. Governments and live stream broadcasters look for ways to detect violent con-

tent before shared publicly on TV channels and other digital and print media. This work targets the detection of fight scenes from social media platforms, in particular, Twitter. Prior solutions for fight detection mostly utilize temporal information from video sequences. However, benefiting from temporal information is not possible to recognize fight actions in still images which are abundant on social media. Thus, the previous approaches cannot be directly adapted for the given task. Existing fight recognition datasets offer limited context such as crowd violence [14], ice hockey games [24], movies [24, 7], or CCTV recordings [2, 26, 5] and are not useful to detect fight scenes in the wild, i.e., social media. To overcome this limitation, we collect a diverse dataset from social media, named Social Media Fight Images (SMFI), which comprises real-world fight scenes shared by the public using their mobile devices. The shots considered as fight scenes are those in which two or more people are using their bodies or objects with an intention to harm each other physically. Other human interactions such as hugging, falling, throwing an object, e.g., ball, are considered as non-violent scenes and included in the non-fight class as negative samples. This helps to prevent the classification from being biased by other characteristics of images such as background or motion blur. In addition, our data collection rely on keywords in multiple languages allowing to collect images from all around the world. Hence, the fight images in SMFI dataset are closer to real-world scenarios and vary along various dimensions such as gender, race, skin color, fight place, etc. The final dataset, including both fight and non-fight images and video frames from Twitter and Google, consists of a total of 5,691 samples. The dataset is available on GitHub.

Next, various image classification networks are re-tuned on the newly created dataset for a binary classification task between fight and non-fight classes. The classification

¹<https://github.com/sayibet/SMFI>

is exclusively based on the images and their labels, instead of sequences. Hence, the previous methods mostly make use of other inputs such as human detection, pose estimation, or the temporal information in videos. These works adapt object detection results. Considering the nature of ght actions, ambiguous scene characteristics already make it impossible to extract pose or human information from the image using pre-trained networks. Similarly, detecting objects in the scene does not help with ght action recognition since the action itself is not related to certain objects.

Fight recognition on images can also be applied on fashion [36].

video-based datasets as a frame-based classification approach. This enables us to conduct comparative experiments and assess the contribution of temporal information for learning a representation of ght actions on video sequences using the public ght datasets [14, 24]. We observe that these datasets are likely to be simpler datasets where classification of videos is possible even using a random frame extracted from each video. To this end, the proposed SMFI dataset is compared with the publicly available video ght datasets in terms of their generalization abilities through cross-dataset experiments.

The main contributions of the study can be summarized as follows:

- We present the Social Media Fight Images (SMFI) dataset that contains a diverse range of ght scenes captured in the wild.
- We show that the ght actions can be detected successfully from still images. We are able to reach 95% accuracy on the collected-in-the-wild dataset.
- Through cross-dataset experiments, we show that the dataset bias issue exists also for ght recognition problem. The results also indicate that SMFI dataset is one of the two most representative datasets among the utilized ve ght datasets.

In the remainder of the paper, we review and discuss the related work in Section 2, describe our methodology in Section 3, present details of the proposed dataset in Section 4, explain the experimental setup, share and analyze the results in Section 5, and conclude the paper in Section 6.

2. Related Work

Previous works can be grouped under two subsections as violence and ght detection and action recognition from still images.

2.1. Violence and ght detection

The violence and ght detection problem is tackled in many aspects, and datasets for various use cases are created and published such as sports games and movies [24], crowd violence [14], surveillance cameras [32, 2, 26, 5]. However, all of these datasets are created mainly for detecting anomalies, violence, or specially ghts, in video

There are a few works addressing violence detection on social media using visual features, as well. In [27], authors recognized the videos with violent content by the acceleration information between consecutive frames. Besides, a multi-modal system was presented in [4] where both image and text of the social media posts were processed in order to detect gang violence on social media. Concerning the violence detection in still images, [35] proposed a new dataset collecting the violence images with keywords such as violence, horror, ght, explosion, blood, gun, and classified violence and non-violence images using bag-of-words. Different than this work, we focus primarily on ght scenarios to learn a more specific representation. Besides, our SMFI dataset is larger than the dataset proposed in [35] which has 500 violence and 1500 non-violence samples.

2.2. Action recognition from still images

Aforementioned violence detection methods are generally applied on video samples. However, the introduced problem aims to recognize ght scenes from single images without using any temporal information, which falls under the area of action recognition from still images. The earlier works in this field predominantly utilize pose-oriented or context-based approaches. [16] used hand-crafted pose features extracted from human performing the act. Similarly [23] and [28] employed a pose estimation network to learn actions from poses. Color information on images were also used for action recognition as proposed in [19]. Assuming that there is a strong relation between the objects in the scene and the performed action, human-object interactions were exploited in [6]. Besides, the context of the scene such as objects or the environment around the performer was seen as an informative clue and used by [12]. Person detector [12] or human part detector [29] networks were also attached to the pipelines in order to keep the focus on the target human in the scene and the performer's pose. Most of these works require additional input such as human bounding boxes, object annotations, etc. rather than using the image-level action labels as is. Therefore, in [41], authors proposed a system that recognized the actions solely based on the image labels without any additional annotations by predicting the human-object interactions during the training. Similarly, in [21], human-mask loss was proposed which directed the activation on feature maps to

the person in action automatically. Context information was also extracted using region proposals and pyramid network as stated in [39]. [40] used ensemble deep neural networks to learn actions from images. [3] detected the salient areas on images and then used multi-attention networks to recognize actions with the help of salient points. [37] proposed two modules for capturing human-object and scene-object relations on action images.

Nevertheless, for the case of ght recognition in the wild, it is not convenient to get an additional clue from surrounding objects or the context of the scene. Besides, because ghting is an abnormal activity, poses of the people cannot be properly extracted from a single image or person detection does not perform well due to unusual poses and occlusions. Having a similar perspective, in [22] CNN models were re-tuned to recognize actions from web images and an extensive experimental analysis is conducted on the ability of CNNs to recognize actions from still images with no additional input.

3. Methodology

The task of ght detection from social media images is a binary classification problem, where the scenes including ght actions should be discriminated from non-ght scenes. For image-based classification, Convolutional Neural Networks (CNN) have been widely used and their ability for image classification is already proven. Recently, Vision Transformer (ViT) [9] has drawn interest with a promising performance on visual tasks. Hence, we also evaluated the performance of ViT on ght recognition from still images problem. Using ViT, we benefited from the self-attention mechanism for learning the alignment between different regions of an image. For our case, learning this alignment gives valuable information regarding the relevant positions of people and their body parts, which is significant for ght recognition. In addition, it is shown in [25] that ViT is robust to blurry images which is a common case for the ght images due to motion.

Specifically, ViT segments the images into patches and extracts the patch + position embeddings where the position indicates the location of the patch on the original image. Then these vectors are fed into a transformer autoencoder together with an extra class token. At the output of the transformer autoencoder, a multi-layer perceptron (MLP) takes the class token as the input and predicts the class of given sample. The network is initialized with pre-trained weights and the MLP layer is replaced with a two-class output MLP layer for training it on ght recognition task.

ViT is released with various versions in terms of the size of patches (16/16, 32/32) and the depth of the transformer autoencoder (Base, Large, Huge). For ght recognition, Large ViT with 16/16 patch size is employed which is deep enough to generalize to the task at hand and computational

4. Dataset

The proposed Social Media Fight Images (SMFI) dataset consists of various sample images and video frames collected from Twitter, Google, and NTU CCTV-Fights Dataset [26]. The NTU CCTV-Fights Dataset both includes surveillance camera recordings and mobile camera recordings of ght scenarios. Given that the main objective is to recognize ght scenarios on social media content, we retrieved the video frames from NTU CCTV-Fights Dataset recorded with mobile cameras which are likely to be shared on social media. For the remaining part of the dataset, possible tags and keywords related to ght scenarios in multiple languages are used for crawling images and videos from Twitter and Google.

Figure 1. Distribution of the samples across sources where inner circle indicates the overall distribution across sources and outer circle displays the percentages of classes for each type of source. Overall class percentages are 48.1% for ght class and 51.9% for non-ght class.

Twitter data constitutes the largest part of the dataset with 86% as it can be seen from Figure 1. The media from Twitter have been collected gradually and at each step, gathered media items labeled as ght and non-ght. As a massive amount of the collected media is unrelated to the ght

	Twitter	Google	CCTV [26]	Total
Fight	2247	162	330	2739
Non- ght	2642	146	164	2952
Total	4889	308	494	5691

Table 1. Number of ght and non- ght samples across different sources in the SMFI dataset.

scenarios, the non- ght samples also chosen from this part of the collection. After the rst batch of images and video frames are labeled manually, an initial classi cation model is trained on the rst batch. Then, this model is employed to assign weak labels for the next batch, which were still manually veri ed and corrected if necessary, making the overall labeling task easier.

Several keywords were used for searching ght scenarios on Twitter as ght, school ght, street ght, ghting people among others. We used the publicly available AIDR system for data collection on Twitter [17]. Furthermore, we also considered that the social media updates are mostly regional and the keywords also depend on the language of the user. Consequently, the set of search keywords are extended including multiple languages as French, Chinese, Russian, Arabic, Spanish, Hindi, Turkish and such. Searching keywords in different languages provided us a more diverse set of images displaying ght actions with various individuals from all around the world so that the model would not be biased towards any particular race or geographic region. Additionally, ght positions (i.e., kicking, wrestling on the ground, punching) of the individuals and number of individuals on the scene vary across the samples.

As the proposed SMFI dataset is in-the-wild dataset which comprises recordings of ght moments in real world and the recognition domain is social media content, the non-ght samples of the dataset are also collected from Twitter. There are different types of non- ght instances as easy, normal, and hard samples. Easy samples are the images that are totally unrelated to the real world such as screenshots, memes, and similar content that are likely to be shared on social media. Normal samples are selfies, real world photographs without people in them or with people standing still. Hard samples are such images or video frames where the people in the scene are mostly the samples which were misinterpreted by the initial model. The sport videos with players running or throwing ball, dancing people, some crowded scenes, and blurry images can be categorized under hard samples. As many as possible hard samples are included in the dataset so that the classi cation will be solely based on the action displayed on the images instead of any other characteristics such as number of people in the scene or motion blur. The number of samples in the SMFI dataset is given in Table 1.

5. Experiments and Results

We performed extensive experiments on multiple datasets including the proposed SMFI dataset. We mainly investigated four research problems: (1) How well can one perform ght recognition from still images in the wild? (2) As the available data on social media changes over time, how does this affect the performance of the trained model? (3) Can a model trained on still images be used for ght recognition on videos? (4) How well do the trained models generalize across different ght datasets?

5.1. Fight recognition on social media images

Fight recognition on still images using the proposed SMFI dataset is investigated in this section. For comparison, various image classi cation networks, such as VGG-16 [30], ResNet-50 [15], ResNeXt-50 [38], and ViT [9] were employed as these networks cover the essential concepts of image classi cation task. We measured the performances of the networks using 10-fold cross-validation rather than using a xed test set. Considering that the samples from the test set might be removed in time, reporting results on a xed test set would hamper the reproducibility of the experimental evaluations. Instead, 10-fold cross-validation could better represent the overall results.

All networks were trained for 20 epochs, and the implementation details for the employed classi cation networks are as follows:

VGG-16: Cross-entropy loss with Adam optimizer was used. Weight decay was 1e-3 and learning rate was 5e-4 for the pre-trained layers and 1e-3 for the nal classi cation layer. First three layers were frozen.

ResNet-50 and ResNeXt-50: Cross-entropy loss with Adam optimizer was used. Weight decay was 1e-3 and learning rate was 5e-4 for the pre-trained layers and 1e-2 for the nal classi cation layer. First ve layers were frozen.

ViT-Large-16: Cross-entropy loss with SGD optimizer was used. Weight decay was 1e-2 and learning rate was 3e-3.

5.1.1 Results

Images in the proposed SMFI dataset were used in the 10-fold cross-validation experiments and resulting average accuracies are reported in Table 2. The results suggest that ViT is superior to other networks for this task by surpassing their validation accuracies with a large gap. This observation demonstrates the generalization ability of ViT as the effect of the over tting is much less than the other networks which are heavily over tting. Qualitatively, ViT successfully learns about the ght action by attending the correct regions of the image as illustrated in Figure 2.

Architecture	Train	Validation
VGG-16	96.3%	83.0%
ResNet50	100%	87.7%
ResNext50	100%	88.3%
ViT Large 16	96.3%	95.5%

Table 2. 10-fold cross-validation results of image classification networks on proposed SMFI dataset. ViT outperforms other network in terms of both validation accuracy and less overfit.

the users themselves. The removal speed of the violent content is relatively higher due to the sensitivity of the subject matter. Eventually, it is unfeasible to retrieve the entire dataset completely as time passes and the number of samples that can be accessed through the shared links is likely to decrease. Therefore, aside from the experiments that uses the whole dataset, we also aimed to observe the effect of removal of data on the performance of the trained model. To that end, the dataset is split into training, validation, and test sets as 70%, 20%, and 10%, respectively. Five experiments were held on five partitions of the dataset as using 40%, 50%, 60%, 80%, and 100% of all training and validation samples while the test set is kept constant for comparability of the results. The removed samples were chosen randomly in order to maintain a consistent distribution across different splits. This experimental setup aimed at simulating the dataset size at different timestamps and investigated the change in the performance of the model with respect to dataset size.

5.2.1 Results

As the surpassing model of the previous section, only ViT is evaluated in this experiment, and the results in Table 3 indicate that the trained ViT model is robust to variations in dataset size. Specifically, when all the samples in the dataset contributed to the learning, the model achieved 95% accuracy on the test split. Considering the training and validation accuracies for the same setup, it can be seen that the model generalized well and showed a decent performance on a relatively hard task. Furthermore, regarding the data loss due to deleted media on social media platforms, even when 60% of the dataset was lost, the model reached results on par with the ideal case (i.e., using the entire dataset). Besides, the effect of data size on overfitting can be observed as the gap between training and validation accuracies decreases as there are more data.

Partition	Train	Validation	Test
100%	95%	92%	95%
80%	95%	92%	94%
60%	96%	91%	94.2%
50%	98%	92%	94.2%
40%	98%	90%	94.2%

Figure 2. Visual attention maps of some violent samples from the SMFI dataset. Maps are extracted from ViT and the model can successfully highlight the salient areas.

Table 3. Performance of ViT with respect to use of different amount of development data.

5.2. Effect of varying dataset size on the model

As we observed during the dataset collection process described in Section 4, the violent content shared on social media platforms is gradually deleted by the authorities or available including Hockey, Movie, Crowd Violence and

5.3. Single frame violent recognition on video datasets

As mentioned before, violent recognition has been studied on video data in general and several benchmark datasets are

	Hockey Fight	Movie Fight	Crowd Violence	Surveillance Fight
Spatiotemporal Encoder [13]	96.5%	100%	92.1%	-
ConvLSTM [31]	97.1%	100%	94.5%	-
Flow Gated Network [5]	98%	100%	88.8%	-
FightNet [42]	97%	100%	-	-
3D CNN [33]	96%	99.9%	98%	-
CNN + Bi-LSTM + attention [2]	98%	100%	-	72%
Kang et al. [18]	99.6%	100%	98%	92%
ViT (frame-based)	98%	100%	98%	84.6%
ResNet50 (frame-based)	99%	99.5%	97%	76.6%

Table 4. Results on video ght datasets. Bold ones are obtained without using temporal information. One frame is chosen randomly from each sample and these frames are classified using respective image classification networks.

Surveillance Fight datasets. The methods applied on these datasets are also video-based approaches that utilize temporal information. In order to get an insight regarding the capacity of still-image-based recognition of ght actions, a comparative experiment has been held on four video ght datasets. The methods applied on these datasets are also video-based approaches that utilize temporal information. In order to get an insight regarding the capacity of still-image-based recognition of ght actions, a comparative experiment has been held on four video ght datasets.

Hockey Fight Dataset [24] contains 1000 videos in total as 500 of them are ght and 500 of them are non-ght samples. The videos are 1-2 seconds long cuts from hockey game recordings where players are ghting or just playing the game.

Movie Dataset [24] contains 200 videos in total where 100 of them are ght samples and 100 of them are non-ght samples. Fight samples are collected from sports games (e.g., soccer, boxing etc.) and some Hollywood movies. Non-ght samples include casual events such as walking, waving hands and more. Videos are 1-2 seconds long.

Violent Flows Dataset [14] is more focused on the crowd violence where high number of performers are involved in the violent act. The dataset consists of 246 samples as 123 of them are ght and 123 of them are non-ght samples. Dataset is collected from real-world scenarios such as violent actions on football games, group ghting on street etc. Duration of the videos vary between 1-6 seconds.

Surveillance Fight Dataset [2] includes CCTV recordings of ght and non-ght occasions collected from YouTube. 300 samples are included in the dataset where 150 of them are ght and 150 of them are non-ght videos. Video sequences are 1-3 seconds long.

For the single frame experiments, one frame was sampled randomly from each video in the dataset and frames were labeled as the label of the video. Two image classification networks are tested on this task as ResNet-50 [15] and ViT [9]. The implementation details are given below. ResNet-50: ImageNet pre-trained network was used where the first few layers were frozen. Learning rate and weight decay were set to 1e-3 and cross-entropy loss with Adam optimizer was used.

5.3.1 Results

Considering the results displayed at Table 4, even if only a single frame was used for classification of videos along with a basic CNN network, it is possible to achieve comparable or sometimes even better performance than the methods that use temporal information. One of the reasons for this result is the inter-class distribution difference between the ght and non-ght classes. Figures 3 and 4 show an overall visual comparison between classes of Hockey and Movie Fight datasets, respectively. For the Hockey Fight Dataset, the recording style differs across the classes as non-ght got in a ght, the camera zoomed in. Similarly for Movie Fight Dataset, the non-ght samples were collected within somewhat controlled environment and the distribution was not the same with the ght occasions collected from movies and sports games. Even the color scale looks discriminative between classes. The mentioned characteristics of these datasets explain why we were able to obtain nearly perfect accuracies without using temporal features at all.

For the Surveillance Fight dataset, even if the image classification networks perform better than CNN + Bi-LSTM + attention model, temporal information can still contribute a lot as the proposed solution by [18] outperforms other methods. Kang et al. [18] proposed two modules for each spatial and temporal attention, which highlight the informative regions in both dimensions. Getting better results with addition of temporal attention indicates that the Surveillance Fight is a relatively harder dataset which may not be classified as successfully by only using spatial information.

Figure 3. Fight class (top) and non-ght class (bottom) frames from Hockey Fight Dataset.

Figure 4. Fight class (top) and non-ght class (bottom) frames from Movie Fight Dataset.

5.4. Cross-dataset Experiments

5.4.1 Results

Cross-dataset experiments were conducted in order to have an insight regarding the trained models' generalizability. For the Hockey, Movie, Crowd Violence, and Surveillance datasets, one frame for each video was used as the testing set. As trained model for video datasets, ResNet-50 model was used for Hockey, Movie and Crowd Violence datasets, and ViT model was used for Surveillance Fight dataset since these models yielded better results at cross-dataset experiments. When testing the models on the proposed SMFI dataset with 10-fold cross-validation, the model trained on the proposed SMFI dataset is able to generalize better than the other three frame-based models trained on video datasets. This explicitly shows that the SMFI dataset spans a wide range of real-world ght recognition scenarios and generalizes well for the problem at hand. Relatively lower accuracies for video-based datasets might mean that the frames extracted from these datasets (Hockey, Movie, Crowd Violence) fail to represent the real-world ght scenarios extensively. It is worth to note that the generalization of Surveillance Fight is also impressive as its average score slightly surpasses the average score of the SMFI dataset.

		Testing				
		Hockey Fight	Movie Fight	Crowd Violence	Surveillance Fight	SMFIAverage
Training	Hockey Fight	-	66.5%	56.5%	62.6%	57.2% 60.7%
	Movie Fight	60.7%	-	60.0%	54.0%	52.2% 56.7%
	Crowd Violence	50.3%	32.0%	-	54.3%	56.8% 48.3%
	Surveillance Fight	77.5%	69.0%	81.4%	-	69.4% 74.3%
	SMFI	70.6%	74.1%	76.7%	67.3%	- 72.2%

Table 5. Cross-dataset experiment results on video-based ght recognition datasets and proposed SMFI dataset. Rows indicate the training dataset and columns indicate the testing dataset.

6. Conclusion

Given the fact that ght recognition is an action recognition problem, existence of temporal information is accepted as an essential part of the previously proposed solutions. However, temporal information is not available all the time and lots of violent media content are shared in image form in social media. This brings the necessity of recognizing ght actions from still images. Nonetheless, the datasets concerning the ght recognition problem are all video-based with limited context and variety. Consequently, we proposed a new dataset named Social Media Fight Images (SMFI) where the samples are collected from social media and mobile camera recordings. Instead of using ght scenarios demonstrated in controlled environments, the real-world spontaneous ght actions are chosen with the intention of having aim-the-wild dataset. We have shown that images containing ght and non-ght actions can be differentiated with a high accuracy even just using still images. Besides, the effect of the removal of data is simulated as well, since the dataset size may not be consistent over time due to deleted social media images. The experimental results indicated that the trained model is robust to changes in the dataset size and can produce stable results even when 60% of the dataset is absent.

Further experiments on video-based ght recognition datasets show that the classification of these datasets can be done successfully using only spatial information from the randomly chosen frames. In addition, the models trained on the proposed dataset and on the video-based datasets are compared via cross-dataset experiments. The results pointed out that the proposed dataset is one of the two most representative datasets among the utilized ght datasets, leading to higher accuracies on unseen datasets.

References

- [1] Simone Accattoli, Paolo Sernani, Nicola Falcionelli, Dagmawi Neway Mekuria, and Aldo Franco Dragoni. Violence detection in videos by combining 3D convolutional neural networks and support vector machines. *Applied Artificial Intelligence* 34(4):329–344, 2020.
- [2] Şeymanur Akt, Özde Ayşe Tatavul, and Hazım Kemal Ekenel. Vision-based ght detection from surveillance cameras. In *2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6, 2019.
- [3] Seyed Sajad Ashra, Shahriar B Shokouhi, and Ahmad Ayatollahi. Action recognition in still images using a multi-attention guided network with weakly supervised saliency detection. *Multimedia Tools and Applications*, pages 1–27, 2021.
- [4] Philipp Blandfort, Desmond U Patton, William R Frey, Sverbor Karaman, Surabhi Bhargava, Fei-Tzin Lee, Siddharth Varia, Chris Kedzie, Michael B Gaskell, Rossano Schifanella, et al. Multimodal social media analysis for gang violence prevention. *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 114–124, 2019.
- [5] Ming Cheng, Kunjing Cai, and Ming Li. RWF-2000: An open large scale video database for violence detection. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4183–4190. IEEE, 2021.
- [6] Vincent Delaitre, Josef Sivic, and Ivan Laptev. Learning person-object interactions for action recognition in still images. In *NIPS 2011: Twenty-Fifth Annual Conference on Neural Information Processing Systems*, 2011.
- [7] Claire-Hélène Demarty, Edric Penet, Mohammad Soleymani, and Guillaume Gravier. VSD, a public dataset for the detection of violent scenes in movies: Design, annotation, analysis and evaluation. *Multimedia Tools and Applications* 74(17):7379–7404, 2015.
- [8] Chunhui Ding, Shouke Fan, Ming Zhu, Weiguo Feng, and Baozhi Jia. Violence detection in video by using 3D convolutional neural networks. *International Symposium on Visual Computing*, pages 551–558. Springer, 2014.
- [9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [10] E Fenil, Gunasekaran Manogaran, GN Vivekananda, T Thanjaivadivel, S Jeeva, A Ahilan, et al. Real time violence detection framework for football stadium comprising of big

- data analysis and deep learning through bidirectional LSTM. *Computer Networks* 51:191–200, 2019.
- [11] Yuan Gao, Hong Liu, Xiaohu Sun, Can Wang, and Yi Liu. Violence detection using oriented violent ows. *Image and Vision Computing* 48:37–41, 2016.
 - [12] Georgia Gkioxari, Ross Girshick, and Jitendra Malik. Contextual action recognition with R²CNN. In *Proceedings of the IEEE International Conference on Computer Vision* pages 1080–1088, 2015.
 - [13] Alex Hanson, Koutilya Pnvr, Sanjukta Krishnagopal, and Larry Davis. Bidirectional convolutional LSTM for the detection of violence in videos. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 405–418, 2018.
 - [14] Tal Hassner, Yossi Itcher, and Orit Kliper-Gross. Violent ows: Real-time detection of violent crowd behavior. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–6, 2012.
 - [15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.
 - [16] Nazli Ikizler, R. Gokberk Cinbis, Selen Pehlivan, and Pinar Duygulu. Recognizing actions from still images. *2008 19th International Conference on Pattern Recognition*, pages 1–4, 2008.
 - [17] Muhammad Imran, Carlos Castillo, Ji Lucas, Patrick Meier, and Sarah Vieweg. AIDR: Artificial intelligence for disaster response. In *Proceedings of the 23rd international conference on world wide web*, pages 159–162, 2014.
 - [18] Min-seok Kang, Rae-Hong Park, and Hyung-Min Park. Efficient spatio-temporal modeling methods for real-time violence recognition. *IEEE Access*, 2021.
 - [19] Fahad Shahbaz Khan, Rao Muhammad Anwer, Joost Van De Weijer, Andrew D Bagdanov, Antonio M Lopez, and Michael Felsberg. Coloring action recognition in still images. *International Journal of Computer Vision* 105(3):205–221, 2013.
 - [20] Ji Li, Xinghao Jiang, Tanfeng Sun, and Ke Xu. Efficient violence detection using 3D convolutional neural networks. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8. IEEE, 2019.
 - [21] Lu Liu, Robby T Tan, and Shaodi You. Loss guided activation for action recognition in still images. In *Asian Conference on Computer Vision*, pages 152–167. Springer, 2018.
 - [22] Shugao Ma, Sarah Adel Bargal, Jianming Zhang, Leonid Sigal, and Stan Sclaroff. Do less and achieve more: Training CNNs for action recognition utilizing action images from the web. *Pattern Recognition* 68:334–345, 2017.
 - [23] Subhransu Maji, Lubomir Bourdev, and Jitendra Malik. Action recognition from a distributed representation of pose and appearance. In *CVPR 2011*, pages 3177–3184. IEEE, 2011.
 - [24] Enrique Bermejo Nieves, Oscar Deniz Suarez, Gloria Bueno García, and Rahul Sukthankar. Violence detection in video using computer vision techniques. In *International Conference on Computer Analysis of Images and Patterns*, pages 332–339. Springer, 2011.
 - [25] Sayak Paul and Pin-Yu Chen. Vision transformers are robust learners. *arXiv preprint arXiv:2105.07581*, 2021.
 - [26] Mauricio Perez, Alex C. Kot, and Anderson Rocha. Detection of real-world ghts in surveillance videos. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2662–2666, 2019.
 - [27] Francisco A Pujol, Higinio Mora, and Maria Luisa Pertegal. A soft computing approach to violence detection in social media for smart cities. *Soft Computing* 24(15):11007–11017, 2020.
 - [28] Tangquan Qi, Yong Xu, Yuhui Quan, Yaodong Wang, and Haibin Ling. Image-based action recognition using hint-enhanced deep neural network. *Neurocomputing* 267:475–488, 2017.
 - [29] Gaurav Sharma, Frédéric Jurie, and Cordelia Schmid. Expanded parts model for semantic description of humans in still images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39(1):87–101, 2016.
 - [30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
 - [31] Swathikiran Sudhakaran and Oswald Lanz. Learning to detect violent videos using convolutional long short-term memory. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2017.
 - [32] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6479–6488, 2018.
 - [33] Fath U Min Ullah, Amin Ullah, Khan Muhammad, Ijaz UI Haq, and Sung Wook Baik. Violence detection using spatiotemporal features with 3D convolutional neural network. *Sensors* 19(11):2472, 2019.
 - [34] Waseem Ullah, Amin Ullah, Ijaz UI Haq, Khan Muhammad, Muhammad Sajjad, and Sung Wook Baik. Cnn features with bi-directional LSTM for real-time anomaly detection in surveillance networks. *Multimedia Tools and Applications* 80(11):16979–16995, 2021.
 - [35] Dong Wang, Zhang Zhang, Wei Wang, Liang Wang, and Tieniu Tan. Baseline results for violence detection in still images. In *2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance*, pages 54–57. IEEE, 2012.
 - [36] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *European Conference on Computer Vision* pages 322–339. Springer, 2020.
 - [37] Wei Wu and Jiale Yu. An improved deep relation network for action recognition in still images. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2450–2454. IEEE, 2021.
 - [38] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference*

on Computer Vision and Pattern Recognition, pages 1492–1500, 2017.

- [39] Shiyang Yan, Jeremy S Smith, and Bailing Zhang. Action recognition from still images based on deep VLAD spatial pyramids. *Signal Processing: Image Communication* 54:118–129, 2017.
- [40] Xiangchun Yu, Zhe Zhang, Lei Wu, Wei Pang, Hechang Chen, Zhezhou Yu, and Bin Li. Deep ensemble learning for human action recognition in still images. *Complexity* 2020, 2020.
- [41] Yu Zhang, Li Cheng, Jianxin Wu, Jianfei Cai, Minh N Do, and Jiangbo Lu. Action recognition in still images with minimum annotation efforts. *IEEE Transactions on Image Processing* 25(11):5479–5490, 2016.
- [42] Peipei Zhou, Qinghai Ding, Haibo Luo, and Xinglin Hou. Violent interaction detection in video based on deep learning. In *Journal of Physics: Conference Series*, volume 844, page 012044. IOP Publishing, 2017.
- [43] Peipei Zhou, Qinghai Ding, Haibo Luo, and Xinglin Hou. Violence detection in surveillance video using low-level features. *PLoS One* 13(10):e0203668, 2018.