ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

FIGHT RECOGNITION FROM STILL IMAGES IN THE WILD

M.Sc. THESIS

Şeymanur AKTI

Department of Computer Engineering

Computer Engineering Programme

JUNE 2022

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

FIGHT RECOGNITION FROM STILL IMAGES IN THE WILD

M.Sc. THESIS

Şeymanur AKTI
(504191539)

Department of Computer Engineering

Computer Engineering Programme

Thesis Advisor: Prof. Dr. Hazım Kemal EKENEL

JUNE 2022

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

DOĞAL ORTAMDA SABİT GÖRÜNTÜLERDEN KAVGA TESPİTİ

YÜKSEK LİSANS TEZİ

Şeymanur AKTI
(504191539)

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

Tez Danışmanı: Prof. Dr. Hazım Kemal EKENEL

HAZİRAN 2022

Şeymanur AKTI, a M.Sc. student of ITU Graduate School with student ID 504191539 successfully defended the thesis entitled "FIGHT RECOGNITION FROM STILL IMAGES IN THE WILD", which she prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor :	**Prof. Dr. Hazım Kemal EKENEL** Istanbul Technical University	.....
Jury Members :	**Assoc. Prof. Dr. İlkay ÖKSÜZ** Istanbul Technical University	.....
	Asst. Prof. Dr. Fatma GÜNEY Koç University	.....
		.....

Date of Submission : **3 June 2022**
Date of Defense : **22 June 2022**

To my family and friends,

FOREWORD

My deepest thanks to my advisor Prof. Dr. Hazım Kemal Ekenel for his support and guidance throughout my thesis and academic studies. I learned a lot from him and he has been a great role model for me. Also thanks to Dr. Ferda Ofli and Dr. Muhammad Imran for their valuable supervision during the development of my thesis.

This work is conducted with the support of Scientific and Technological Research Council of Turkey (TUBITAK) ARDEB 1001 Grant No 121E408 and TUBITAK 2210-A National MSc Scholarship. I am grateful for the supports provided by TUBITAK.

I am thankful to SiMiT Lab members with whom we had a lot of good times and fruitful meetings. My special thanks to my teammate Alperen Kantarcı for being a nice and supportive companion. It was a privilege to work with him.

I was able to complete my Master's journey with the best memories thanks to my lovely friends Gamze Akyol, Ali Eren Çelik, Alperen Kantarcı and Azmi Can Özgen.

Finally, I am grateful to my family and my girls who have always believed in me regardless of the subject. Their encouragement is more than precious to me.

June 2022 Şeymanur AKTI

TABLE OF CONTENTS

ABBREVIATIONS

AIDR	**:** Artificial Intelligence for Disaster Response
Bi-LSTM	**:** Bi-directional Long Short-Term Memory
CNN	**:** Convolutional Neural Network
LSTM	**:** Long Short-Term Memory
MLP	**:** Multi-layer Perceptron
SGD	**:** Stochastic Gradient Descent
SMFI	**:** Social Media Fight Images
ViT	**:** Vision Transformer

LIST OF TABLES

Page

LIST OF FIGURES

Page

FIGHT RECOGNITION FROM STILL IMAGES IN THE WILD

SUMMARY

Violence in general is a sensitive subject and can have a negative impact on both the involved people and witnesses. Fighting is one of the most common types of violence which can be defined as an act where individuals intend to harm each other physically. In daily life, these kinds of situations might not be faced too often, however, the violent content on social media is also a big concern for the users. Since violent acts or fights in particular are considered as an anomaly or intriguing for some, people tend to record these scenes and upload them on their social media accounts. Similarly, news agencies also regard them as newsworthy material in some cases. As a result, fighting scenes become available on social media platforms frequently. Some users may be sensitive to these kinds of media content and children who can be harmed due to the aggressive nature of the fight scenes also uses social media. These facts make it necessary to detect and put limitations on the distribution of violent content on social media.

There are some systems focusing on violence and fight recognition on visual data. However, these works mostly propose methods on different domains for violence such as movies, surveillance cameras, etc., and the social media case remains unexplored. Furthermore, even if most of the fight scenes shared on social media are in video sequences, there is also a non-ignorable amount of image data depicting violent fighting. However, no work tackles the fight recognition from still images instead of videos. Thus, in this thesis, the problem of fight recognition from still images is investigated.

In this scope, first, a novel dataset was collected from social media images which is named Social Media Fight Images (SMFI). The dataset was collected from Twitter and Google images and some frames were included from the video dataset of NTU CCTV-Fights [3]. The fight samples were chosen among the samples which are recorded in uncontrolled environments. In order to crawl a large amount of data, different keywords were used in various languages. The non-fight samples were also chosen among the data crawled from social media in order to keep the domain consistent across the classes. The dataset is made publicly available by sharing the links to the images.

For the classification of the Social Media Fight Images dataset, some image classification methods were applied to the dataset. First, Convolutional Neural Networks (CNN) were employed for the task and their performance was assessed. Then, a recent approach, Vision Transformer (ViT) was exploited for the classification of the fight and non-fight images. The comparison showed that the Vision Transformer gives better results on the dataset achieving a higher accuracy with less overfit. A

further experiment was also held on investigating the effect of varying dataset sizes on the performance of the model. This was seen as necessary as the data shared on social media may be deleted in the future and it is not always possible to retrieve the whole dataset. So, the model was trained on different partitions of the dataset and the results showed that even if using more data is better, the model could still give satisfying performance even in absence of 60% of the dataset.

Upon the successful results on fight recognition on still images problem, another experimental study was conducted on the classification of video-based datasets using a single frame from each sample. The experiment included four video-based fight datasets and results showed that three of them could be successfully classified without using any temporal information. This indicated that there might be a dataset bias for these three datasets where the inter-class visual difference is high across the classes. Cross-dataset experiments also supported this hypothesis where the trained models on these video datasets perform poorly on the other fight recognition datasets. Nonetheless, the network trained on the proposed SMFI dataset gave a promising accuracy on other datasets as well, showing that the dataset generalizes the fight recognition problem better than the others.

DOĞAL ORTAMDA SABİT GÖRÜNTÜLERDEN KAVGA TESPİTİ

ÖZET

Şiddet kavramı çoğunluk tarafından hassas bir konu olarak görülür ve olaya karışan veya tanık olan kişiler üzerinde olumsuz etkilere sebep olması kaçınılmazdır. Şiddetin bir türevi olan kavga, kişilerin birbirlerine fiziksel anlamda zarar vermek amacıyla girdikleri etkileşim olarak tanımlanabilir. Günlük hayatta bu tür durumlarla sık karşılaşılmasa da sosyal medya mecralarında paylaşılan şiddet ve kavga içeriklerinin kullanıcılar üzerindeki etkisi bir sorun olarak görülmektedir. Kavga anları olağandışı bir durum olarak görüldüğünden ve ilgi çekebileceği düşünüldüğünden kişiler bu tür olaylara tanık olduklarında olayı kayda alıp sosyal medya hesaplarında paylaşmaya eğilimlidirler. Dahası, medya kurumları da bu tür içeriklerde haber değeri gördüklerinden içeriğin paylaşımı ve yayılmasında etkili olabilirler. Sonuç olarak sosyal medya platformlarında şiddet ve kavga içerikli görüntülere maruz kalmak olasıdır. Bazı kullanıcılar bu tür içeriklere karşı hassas olabilir, benzer şekilde sosyal medya kullanan genç yaştaki bireyler bu tür saldırgan içeriklerden olumsuz etkilenebilir. Bu durumlar göz önünde bulundurulduğunda sosyal medyada şiddet ve kavga içerikli görüntülerin tespit edilerek paylaşımına bir sınırlandırma getirilmesi veya sansürlenmesi ihtiyacı doğmuştur.

Görsel veri üzerinde şiddet ve kavga tespiti üzerine geliştirilen çeşitli sistemler bulunmaktadır. Ancak bu çalışmalar sosyal medya verisi üzerine değil, güvenlik kameraları, filmler veya spor oyunları gibi farklı kavramlar üzerine odaklanmıştır. Ayrıca, sosyal medyada karşılaşılan şiddet içerikli görüntülerin çoğunluğu video formunda olsa da sabit görüntü olarak paylaşılan veri miktarı da ihmal edilemez düzeydedir. Ancak önceki çalışmaların tümü video verisi üzerinde şiddet ve kavga tespiti konusunu ele almış ve sabit görüntüleri ele alan herhangi bir çalışma yapılmamıştır. Video verisinde hareket tanımlama için elverişli olan zamansal bilgiye ulaşmak mümkün iken sabit görüntülerde yalnızca uzamsal bilgiden, yani görüntü üzerindeki piksellerin ilişkisinden yararlanılabilir. Bu durum, özellikle kavga tespiti problemini zorlaştırmaktadır çünkü kavga tespiti de bir nevi eylem tanıma problemidir ve zamansal bilginin önemi bu tür problemler için yüksektir. Bu sebeple bu tezde göreceli olarak zor bir problem olan sosyal medyada sabit görüntülerden kavga tespiti üzerine çalışılmıştır.

Bu kapsamda öncelikle probleme uygun bir veri kümesi toplanmıştır. Önceki çalışmalarda sunulan veri kümelerinin tamamı video temelli veri kümeleridir ve sosyal medya alanında değillerdir. Bu sebeple probleme yönelik Sosyal Medya Kavga Resimleri (SMFI) isimli bir veri kümesi oluşturulmuştur. Veri kümesine dahil edilen görüntüler Twitter, Google gibi sosyal medya mecralarından ve NTU CCTV-Fight [3] isimli video veri kümesindeki video karelerinden toplanmıştır. Veri kümesi "kavga"

ve "kavga olmayan" şeklinde iki sınıf içermektedir. Oluşturulan veri kümesi toplamda 5,691 görüntü içermektedir ve veri kümesi örneklerin bağlantı bilgileri ile kullanıma açık olarak paylaşılmıştır.

Sosyal medyadan toplanan kavga sahneleri kontrolsüz ortamlarda kaydedilmiş görüntülerden oluşmaktadır. Kavga sahnelerinin ani bir şekilde ortaya çıktığı göz önüne alındığında çekimlerin birçoğu etrafta bulunan ve olaya tanık olan kişilerin telefon kameralarından çekilen görüntülerdir. Doğal ortamda kavga tespitine uygun bir veri kümesi geliştirmek için seçilen resimler de bu duruma uygun olarak seçilmiştir. Kontrollü ortamda kaydedilmiş görüntüler ve güvenlik kamerası gibi farklı türlerdeki görüntüler veri kümesine dahil edilmemiştir. Veri kümesini olabildiğince çeşitlendirebilmek için Twitter ve Google'dan veri toplarken çeşitli arama terimleri kullanılmıştır. Sokak kavgası, okulda kavga, çete kavgası gibi arama terimleri kullanılarak olayın gerçekleştiği ortamın ve ortamda bulunan kişi sayısının değişiklik gösterebileceği senaryolar da ele alınmıştır. Ayrıca, yerel sosyal medya paylaşımları ele alındığında, kullanıcılar içeriklerini çoğunlukla kendi ana dillerinde paylaşma eğilimindedir. Farklı coğrafik bölgelerden veri toplayabilmek için arama terimleri kümesi farklı diller de eklenerek genişletilmiştir. Böylece toplanan veri kümesine farklı milletlerden insanlar içeren örnekler de dahil edilmiş ve veri kümesinin taraflı olmasının önüne geçilmiştir.

Veri kümesinin kavga olmayan sınıfı da bağlama sadık kalınarak sosyal medya görüntülerinden toplanmıştır. Kavga olmayan sınıfta farklı zorluk seviyelerinde görüntüler dahil edilmiştir. Ayırt edilmesi daha kolay olan örnekler, sosyal medyada paylaşılan ve kontrolsüz ortamlardan uzak, karikatürler, ekran görüntüleri, özçekimler vs. olarak tanımlanabilir. Orta düzeydeki kavga olmayan örnekler gerçek dünyada çekilmiş dış mekan / iç mekan görüntüleri ve normal davranışlar sergileyen insanları içeren bir takım görüntülerdir. Ayırt edilmesi zor olarak görülen örnekler de kavga ile karıştırılabilecek davranışları içeren örneklerdir ve bunlar sarılan insanlar, spor oyunlarında yarışan insanlar, bir şey fırlatan insanlar vb. olarak tanımlanabilir.

SMFI veri kümesinin sınıflandırılması için çeşitli görüntü sınıflandırma yöntemleri test edilmiştir. Sınıflandırma aşaması için görüntülerde insan tespiti ve poz belirleme gibi bilgilerden yararlanılması planlansa da kavga gibi anormal bir davranışın bu tür bilgilerin çıkarılması için uygun olmadığı görülmüştür. Görüntüde üst üste gelen nesneler ve insanlar, olağandışı pozisyonlar ve gerçek ortamda çekilen düşük çözünürlüklü görüntüler bu tür ek bilgilerin kullanımına imkan sağlamamıştır. Bu sebeple yalnızca RGB resimler kullanılarak sınıflandırma sağlanmıştır. Öncelikle, görüntü temelli problemlerde yaygın bir şekilde kullanılan Evrişimli Sinir Ağları'nın (Convolutional Neural Network) çeşitli türevleri veri kümesi üzerinde eğitilmiş ve test edilmiştir. Evrişimli Sinir Ağları istenen sonucu veremediği için daha güncel bir yöntem olan Görsel Dönüştürücü (Vision Transformer) ağı kavga ve kavga olmayan sahnelerin sınıflandırılması için kullanılmıştır. Görsel Dönüştürücü yapısındaki öz-dikkat (self-attention) mekanizması sayesinde modelin görüntüden faydalı bilgiler çıkararak başarılı bir şekilde sınıflandırma yaptığı görülmüştür. Bu başarının bir diğer sebebi de hareket sebebiyle bulanık olabilen kavga görüntülerinin bulanık görüntülere karşı dirençli olan Görsel Dönüştürücü ile başarılı bir şekilde tanınması olarak görülmüştür.

Veri kümesinin tamamı üzerinde yapılan deneylerin yanı sıra, modelin değişen veri kümesi boyutundan ne şekilde etkilendiği üzerine de bir deney yürütülmüştür. Sosyal medyada paylaşılan veriler zamanlar silinebileceğinden gelecekte veri kümesinin tamamına erişimin mümkün olmayabileceği göz önüne alındığında modelin bu değişiklikten nasıl etkileneceğinin incelenmesi de önem arz etmektedir. Bunun için görüntü sınıflandırma modeli veri kümesinin farklı miktardaki parçaları üzerinde eğitilmiş ve test edilmiştir. Sonuçta her ne kadar veri azaldıkça doğruluk oranı düşüyor olsa da, veri kümesinin %60'lık kısmının kaybedildiği senaryoda bile modelin başarılı bir şekilde sınıflandırma yapabildiği görülmüştür.

Sabit görüntülerde kavga tespiti çalışması sonucunda elde edilen başarılı sonuçlar ışığında video temelli veri kümelerinin de sabit görüntüler üzerinden sınıflandırılıp sınıflandırılamayacağı konusu araştırılmıştır. Bu durumu incelemek için dört adet video temelli veri kümesi seçilmiştir, bunlar Hockey, Movie, Crowd Violence ve Surveillance Fight veri kümeleridir. Bu veri kümelerindeki her bir video örneğinden rastgele bir kare seçilmiş ve bu kareler veri kümelerinin sınıflandırılması için kullanılmıştır. Deneyde video temelli kavga tespiti yöntemleri ile tek kare üzerinde eğitilmiş görüntü sınıflandırma yöntemleri karşılaştırılmış ve ilgi çekici sonuçlar elde edilmiştir. Video temelli veri kümelerinin üçünde (Hockey, Movie ve Crowd Violence) zamansal bilgiden yararlanmadan eğitilen görüntü sınıflandırma modellerinin video temelli yaklaşımların başarımına ulaştığı görülmüştür. Bu veri kümelerindeki videolar incelendiğinde sınıflar arasındaki görsel farkın yüksek olduğu görülmüştür. Bu sebeple yalnızca tek bir kareye bakarak örneğin ait olduğu sınıf belirlenebilmektedir.

Elde edilen bu sonuçlar video temelli kavga veri setlerinde bir yanlılık (bias) problemi olması ihtimaline işaret etmektedir. Bu konuyu daha detaylı inceleyebilmek adına çapraz veri kümesi deneyleri yapılmış ve her bir veri kümesi üzerinde eğitilen model diğer veri kümeleri üzerinde test edilerek doğruluk oranlarının ortalaması hesaplanmıştır. Bu deneyde elde edilen sonuçlar Hockey, Movie ve Crowd Violence veri kümelerinde eğitilen modellerin aynı veri kümelerinde test edildiğinde mükemmele yakın sonuç vermelerine rağmen, farklı bir kavga veri kümesi üzerinde test edildiğinde oldukça düşük performans elde edildiğini göstermiştir. Bu sonuç, bahsi geçen veri kümelerinin kavga tespiti problemini genelleme konusunda başarısız olduğunu ve bu veri kümelerinde yanlılık problemi olduğunu göstermiştir. Diğer yandan tez kapsamında geliştirilen SMFI veri kümesi üzerinde eğitilen model, diğer veri kümeleri üzerinde de tatmin edici bir performans sergilemiştir ve SMFI veri kümesinin kavga tespiti için yeteri kadar kapsamlı bir veri kümesi olduğu gösterilmiştir.

Sonuç olarak tez kapsamında daha önce ele alınmamış olan sabit görüntülerde kavga tespiti konusu ele alınmıştır. Bu kapsamda öncelikle probleme uygun bir veri kümesi oluşturulması için sosyal medya platformlarından kavga görüntüleri toplanmış ve etiketlenmiştir. Bu veri kümesinin sınıflandırılması için görüntü sınıflandırma ağları test edilmiş ve kapsamlı bir deneysel analiz gerçekleştirilmiştir. Elde edilen sonuçlar Görsel Dönüştürücü modelinin sunulan veri kümesini başarıyla sınıflandırabildiğini ve değişen veri kümesi boyutuna karşı dirençli olduğunu göstermiştir. Ayrıca video temelli veri kümeleri üzerinde yapılan karşılaştırmalı deneyler bu veri kümelerindeki

yanlılık (bias) problemini ortaya çıkarmış ve sunulan SMFI veri kümesinin kavga tespiti problemini yeterince iyi bir şekilde kapsadığı görülmüştür.

1. INTRODUCTION

Violence is an invincible yet harmful concept in human life for both the involved people and the witnesses. The negative effect of violence on people is undeniable and it is expected that people tend to avoid witnessing violent situations such as fighting. However, considering the unrestrained nature of the feelings such as anger, injustice, hate, etc. it is not possible to prevent violent acts to be happening all the time. This results in unwanted danger and also unfavorable effects on society. Even if people do not face these kinds of violent situations in daily life often, as social media is heavily used for the exchange of information nowadays, they may be exposed to violent contexts undesirably on various platforms. It is not always possible to avoid these media contents completely and especially younger users might not be able to have control over the information they are exposed to. Yet, harmful visual content can cause severe effects on children's development and mental health. This is the reason why some social media platforms have limitations regarding the distribution of these types of media content. As expected, reviewing all of the content available on social media manually is not possible, thus automatic recognition systems are needed for discriminating and eliminating violent content.

Recognizing the media content including fight actions could be seen as a significant step toward limitation of the violent content on social media since a remarkable part of the violent acts consists of fighting. In addition, the fighting is found to be intriguing to some where people tend to record these moments and share them on their social media accounts. Also in some cases, the press might be involved in reporting these sorts of anomalies and releasing related content on social media. This results in a huge amount of violent content existing on social media platforms. The mass media content regarding the violence is shared in video sequences, however, there is a non-negligible amount of image data on social media as well.

Recognition of violent acts, specifically, the fight action from still images remained an unexplored problem in the literature where most of the works focus on violence or fight detection from video sequences. Considering the amount of still image data on social media platforms, recognizing the violent content in still images is also necessary. The purpose of this thesis is to investigate the capability of deep learning methods to discriminate the image data on social media including fight or not.

In this light, a novel fight image dataset collected from Twitter, Google, and NTU CCTV-Fights [3] is introduced which is named Social Media Fight Images (SMFI). The dataset consists of 5,691 images with extensive representation of fight scenes with varying fight scenarios, backgrounds, the number of people in the scene, and the nationality of the people in the scene. In order to have better coverage, various keywords in various languages were used during data collection. The samples in the fight class were collected from real-life instances of fight actions in uncontrolled environments. Also for the non-fight class, the samples were chosen among the social media content with different difficulty levels. Through these data collection techniques, a non-biased dataset was achieved with high generalizability for the fight recognition task in-the-wild.

Various deep learning architectures are employed for the classification of the fight and non-fight images in SMFI and their performances are compared. One of the concerns regarding the dataset was the removal of data in time due to the deleted posts by users or authorities. This means that the dataset could not be retrieved as a whole in the future. Thus, a second experiment is conducted to investigate the effect of varying dataset sizes on the trained model. The experimental results showed that the trained model is robust to the decreasing dataset size with a satisfying performance even when 60% of the dataset was removed.

Getting promising results on the SMFI dataset, the possibility for classification of the video-based fight datasets using a single frame from each video was investigated as well. For this experiment, the performances of image classification models trained on a single frame from each video were compared to the video-based methods. In the result, the necessity of temporal information for the classification of video-based fight datasets

was examined. Lastly, in order to get a better insight regarding the generalization of fight recognition datasets, cross-dataset experiments are conducted. When the obtained results were compared, the proposed Social Media Fight Images dataset showed its superiority in terms of generalization of the fight recognition task.

Social Media Fight Images dataset and extensive experimental analysis presented in the scope of this thesis provide a significant contribution to the literature as it demonstrates various aspects of the fight recognition problem and a comparative study.

2. LITERATURE REVIEW

There are two main research areas related to fight detection from still images on social media as violence and fight detection and action recognition from still images.

2.1 Violence and Fight Detection

The violence and fight detection problem was investigated in various contexts and several datasets were created and made available to the public. For instance, Movie dataset [4] tackled the problem of recognizing fight actions in movies and sports games and created a dataset concerning this problem. Another dataset was created for detecting crowd violence [5], where authors gathered violent scenes in crowded scenarios, such as football games or group fights on the streets. Recognizing fight actions on surveillance cameras has been seen as an important task by many researchers as it allows to detect the fight actions quickly with help of surveillance cameras located in various places. In this context, [6] introduced an anomaly dataset on surveillance cameras including fight actions together with other anomalies such as abuse, robbery, burglary, and more. More specifically [3,7,8] created datasets focusing on detecting fight actions from surveillance cameras. However, all of the mentioned datasets consist of video sequences of anomalies, violence, or fights. Hence, the previous methods mostly focus on recognizing violence or fight in video sequences by making use of temporal information.

These works adapt existing action recognition methods for violence detection, such as 3D Convolutional Neural Networks for violence detection [9]–[12], recognize violent acts by sequence learning with Recurrent Neural Networks [13]–[16], learn low-level features for violence detection [4,17,18], or even combine visual and audio features in a multi-modal fashion [19].

Some other related works address violence detection on social media using visual features, as well. In [20], acceleration information between consecutive frames is used

for recognizing the videos with violent content. Besides, [21] presents a multi-modal system where both image and text parts of the social media posts were processed in order to detect gang violence on social media.

As for the violence detection in still images, [22] collected a new dataset of images using keywords such as violence, horror, fight, explosion, blood, and gunfire. A classification model for violent and non-violent images in the dataset was developed based on the bag-of-words method. Even if the objective is similar to the task tackled in the scope of this thesis, this thesis focuses primarily on fight scenarios to learn a more specific representation different than [22] which handles a more generic definition of violence on social media. Moreover, the newly collected dataset in the scope of this thesis, Social Media Fight Images (SMFI) is relatively larger than the dataset proposed in [22] which has 500 violence and 1500 non-violence samples.

2.2 Action Recognition from Still Images

The aforementioned violence detection methods were generally applied to video samples. However, the introduced problem aims to recognize fight 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. [23] used hand-crafted pose features extracted from the human performing the act. Similarly [24] and [25] employed a pose estimation network to learn actions from poses. Color information on images was also used for action recognition as proposed in [26]. Assuming that there is a strong relationship between the objects in the scene and the performed action, human-object interactions were exploited in [27]. Besides, the context of the scene, such as objects or the environment around the performer, was seen as an informative cue and used by [28]. Person detector [28] 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 they are. Therefore, in [30], the authors proposed a system that recognizes the actions solely based on the image labels

without any additional annotations by predicting the human-object interactions during the training. Similarly, in [31], the 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 networks as stated in [32]. [33] used ensemble deep neural networks to learn actions from images. [34] detected the salient areas on images and then used multi-attention networks to recognize actions with the help of salient points. [35] proposed two modules for capturing human-object and scene-object relations on action images.

Nevertheless, in the case of fight recognition in-the-wild, it is not convenient to get an additional cue from surrounding objects or the context of the scene. Besides, since fighting is 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 [36] CNN models were fine-tuned to recognize actions from web images, and extensive experimental analysis is conducted on the ability of CNNs to recognize actions from still images with no additional input.

3. METHODOLOGY

In the scope of this thesis, the fight recognition from social media images problem was tackled as a binary classification task and the scenes were classified as including fight action or not. This section summarizes the methodologies which are applied to the problem.

3.1 Convolutional Neural Networks

For image-related tasks such as image classification, object detection, and segmentation, Convolutional Neural Networks (CNN) are widely used where the layered convolution operations make it possible to learn the spatial information on images. Regular Convolutional Neural Network architectures consist of various types of layers such as convolutional layers, pooling layers, fully connected layers, and activation layers. The network aims to take an input image and output a vector representing the prediction scores for each class. Thus, the image is downsampled at convolutional layers, and pooling layers. Convolutional layers apply spatial convolution operation on feature maps with different kernel sizes for learning the spatial relations between pixels. The filter parameters are learned during training. Pooling layers are used for shrinking the size of feature maps as it requires a great number of convolutional layers to obtain a proper feature map size for fully connected layers. In the end, the 2D feature maps are flattened and fully-connected layers are utilized for the last few layers. Fully connected layers take 1D input which contains spatial relation information of images as elements and outputs the prediction scores for each class. Activation layers are used to decide which parts of the map will be activated for the next layer and they add non-linearity to the network. In order to stack multi-layer convolution operations and build a deep network, non-linear activation functions are applied on feature maps after each convolution.

The initial architectures of CNNs, such as LeNet [37] and AlexNet [38], consist of several layers which are considered relatively shallow models. The advancements in computational technology made it possible to train larger networks that can learn from larger datasets. Thus, deeper networks with a larger amount of parameters are also developed for more complex problems such as VGG-16 [39]. Introducing deeper models exposed some problems while training very deep convolutional networks such as vanishing gradients. As the gradient should be passed through many layers in deep networks, after some time, the value for gradients becomes extremely small that the model parameters cannot be updated further. In order to enable training of even deeper models by preventing the vanishing gradient problem, residual blocks are introduced which are utilized in ResNet [40] models.

3.2 Transformers

Aside from the computer vision problems, Transformers [41] has been a superior method for sequence learning and natural language processing. Transformers take a sequence as an input and output another sequence where the self-attention mechanism within the Transformers better represents relations between elements in the sequence and their dependencies. This results in learning more informative representations of sequences for many tasks such as machine translation, language generation, and so on. Self-attention mechanisms were also employed for image-related tasks in order to find better locations on the images to focus on.

3.3 Vision Transformers

More recently, Vision Transformer (ViT) [1] is introduced and has drawn interest with a promising performance on visual tasks. Vision Transformer consists of an embedding layer, transformer encoder, and multi-layer perceptron (MLP) layer. The network regards the patches of an image as a sequence and learns the patch-wise relations using Transformer encoders. 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 the given sample. As the patches are embedded into linear feature vectors, Vision Transformer does not use any convolution operation and still can learn image representations successfully. The overall architecture of the Vision Transformer is given in Figure 3.1.

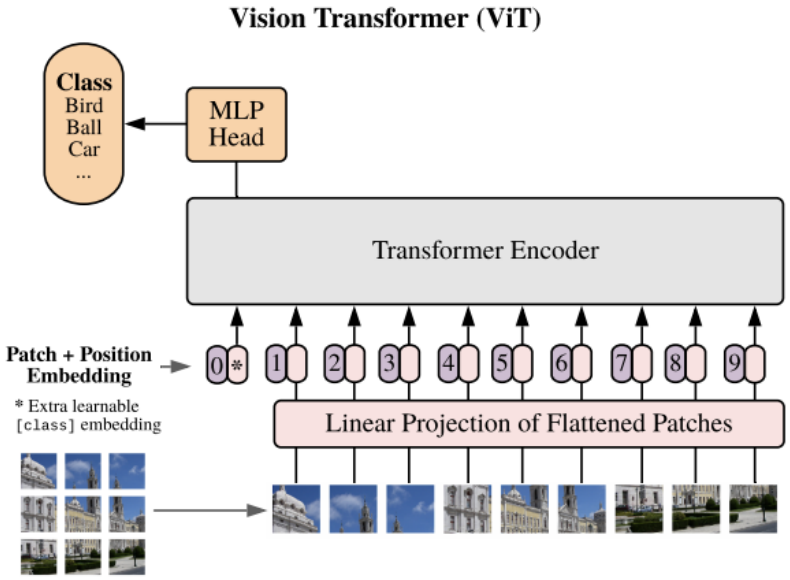

Figure 3.1 : Vision Transformer architecture taken from [1].

With the promising performance of Vision Transformer, the performance of ViT on fight recognition from still images problem was also examined. For the task at hand, learning the alignment between various regions of the fight and non-fight images give valuable information regarding the relevant positions of people and their body parts. This information is seen as a significant cue for fight recognition. One of the concerns for the fight recognition is that the images might contain motion blur due to the aggressive nature of the fight actions. This could hurt the performance of the classifier since some of the classification models might not be able to extract spatial information successfully from blurry images. However, it is shown in [42] that ViT is robust to blurry images where it can still extract valuable information in a challenging setup. These points highlight the advantages of using Vision Transformer for the task at the hand. As discussed in the upcoming experimental results section, the comparison

between the Vision Transformer and the other image classification networks proves its superiority and robustness.

Vision Transformer has various versions released with respect to the size of patches (16×16, 32×32) and the depth of the Transformer autoencoder (Base, Large, Huge). For the fight recognition in still images problem, the Large Vision Transformer with 16×16 patch size was employed. Large Vision Transformer was the most appropriate size for the collected dataset where a model was aimed with proper generalization. The Base Vision Transformer was too shallow for the task and the Huge Vision Transformer is too deep so that it overfits on a relatively small dataset and also computationally expensive. Moreover, optioning patch size as 16×16 enables the network to capture more fine-grained regions and this helps to analyze the images where the fighting people occupy a relatively smaller area within the overall image due to their distance to the camera.

As a common approach in the literature, transfer learning is used for adapting a high-capacity network to a downstream task. Thus, the employed Vision Transformer network weights were initialized as ImageNet pre-trained weights, and the MLP layer at the output of the Transformer autoencoder was replaced with a two-class output MLP layer for fine-tuning it on the binary fight classification task.

For the sake of extracting more information from the fight images, some other modalities were also tried to be employed for the fight recognition problem, such as human segmentation and pose estimation. However, the abnormal positions of the people in the scene and the low-quality images collected from real-world scenarios make it inconvenient to extract human segmentation and pose skeletons properly. Some cases such as people wrestling on the ground and people fighting where one of them is occluded by the other one are examples of inconvenient cases. This is the reason why only RGB images of fight scenes are used for recognition.

4. PROPOSED DATASET

Upon the absence of a dataset to be utilized for fight recognition from still images, a new dataset was generated from the social media images and video frames collected from Twitter, Google, and NTU CCTV-Fights Dataset [3]. The newly generated dataset was named Social Media Fight Images and the image URLs and video frame numbers were shared through Github[1] in order to not violate the copyright of mentioned sources. The distribution of the samples among the sources and classes is given in Figure 4.1.

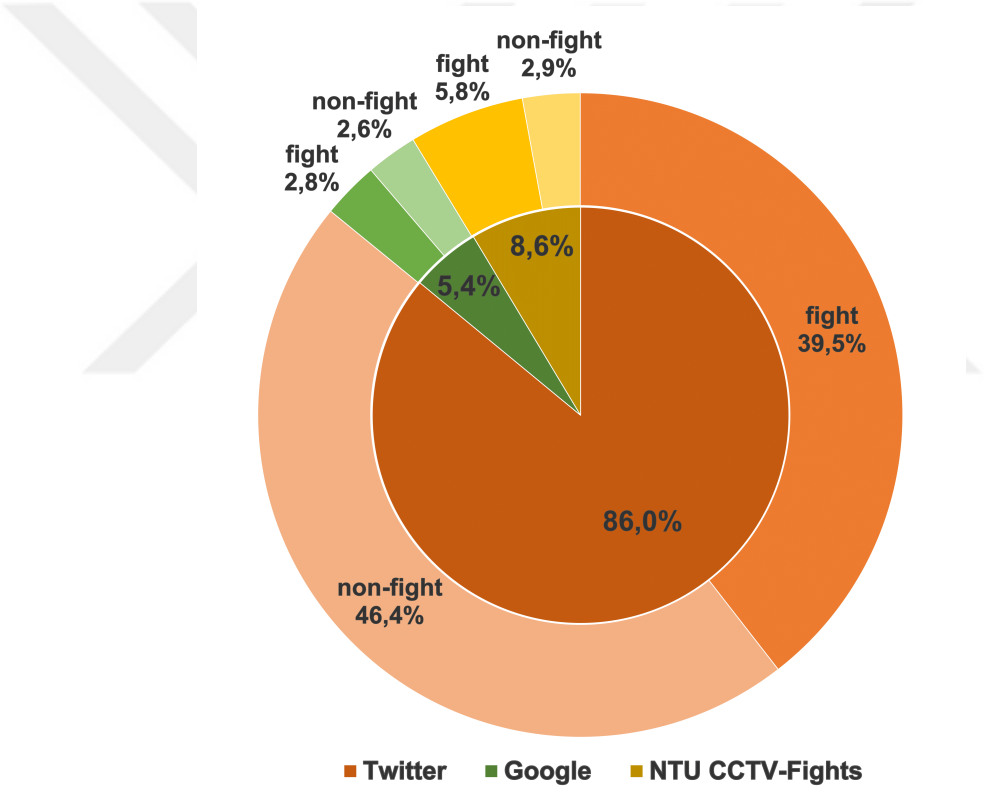

Figure 4.1 : Distribution of the samples across sources where inner circle indicates the overall distribution across sources and outer circle displays the percentages of classes for each type of source. Overall class percentages are 48.1% for fight class and 51.9% for non-fight class.

[1] https://github.com/sayibet/SMFI

As Figure 4.1 shows, Twitter data constitutes the largest part of the dataset with 86% of whole samples. The Twitter data was collected using the publicly available Artificial Intelligence for Disaster Response (AIDR) system [2] in order to gather a large number of images easily. The AIDR system was originally proposed for analyzing tweets during disaster scenarios for assessment of the situation. However, it also aims to introduce a system for easing data collection on the Twitter platform. An overall view of the system is given in Figure 4.2.

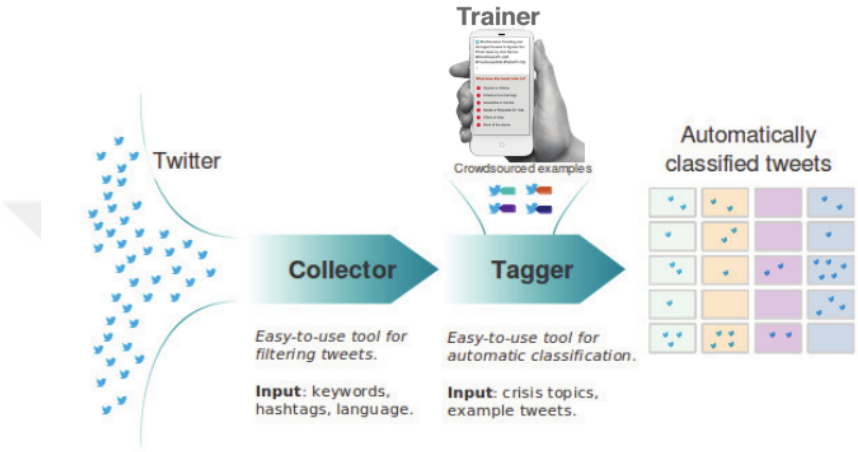

Figure 4.2 : Overall view of AIDR system taken from [2].

The AIDR system first collects tweets from Twitter by filtering them according to the given keywords and hashtags. Then, it uses a crowd-sourcing system to label a subset of collected tweets for training a classifier on these labels. The automatic classification model can be enhanced when more labels are available. In the end, classified tweets are given as the output. For this study, the AIDR system was used for collecting fight-related tweets with media content in them such as images and videos. For the remaining part of the dataset, (1) a script was used for downloading image search results from Google given search keywords, (2) the video frames were manually extracted from NTU CCTV-Fights [3] dataset.

In order to collect a diverse set of images from Twitter and Google, a large set of keywords were used. The keyword set includes but is not limited to *fight, school fight, street fight, fighting people, kicking*. Using a variety of keywords enabled to obtain a generalized dataset in terms of the type of fight action (fisting, kicking, pushing,

wrestling, etc.). It also provided samples that cover different fight scenarios with a varying number of individuals in the scene, such as group fighting, fighting people in crowded scenes, and one-to-one fights. Moreover, considering that the users on social media tend to upload their content in their own language, the tweet texts, hashtags, and keywords may vary based on the user's language and region. In this scope, the keyword set is extended to include keywords in different languages such as French, Chinese, Russian, Arabic, Spanish, Hindi, Turkish. The benefit of this extension was two-fold. First, this made it possible to span a larger part of the social media content and increased the size of the collected data. Second, taking into account that the individuals displayed in the media content may vary in terms of race based on the user's country, collecting samples from all around the world enabled to have a diverse dataset that is not biased towards any particular race or geographic region.

The images from social media platforms were collected based on chosen keywords. Since it is common to share media content on social media irrelevant to the text, keywords, or hashtags, a large part of the images were not relevant to the fight context. Thus, the gathered set of images was examined further by checking their content manually and choosing appropriate samples for fight class. The same process is applied to the NTU CCTV-Fights [3] part of the dataset where frames from videos in the dataset were uniformly sampled and diverse frames were chosen among them. During this examination, for both the social media data and NTU CCTV-Fights [3] dataset frames, some constraints were considered. For instance, people tend to share their own recordings of fighting moments on social media where most of the samples are spontaneous mobile device recordings. Thus, as the main domain is recognizing fight images on social media, samples with these characteristics were chosen. The surveillance camera recordings in NTU CCTV-Fights [3] dataset or controlled recordings of fight actions are not included in the dataset. Besides, the gender, race, and positions of the actors vary across the chosen samples.

The dataset was collected gradually. In order to make it easier to annotate the collected raw data, an intermediate classification model was trained on the first batch of the dataset. The model was then used to assign weak labels for the upcoming batches. The

weak labels were reviewed and corrected if necessary. This made the dataset collection and annotation much easier.

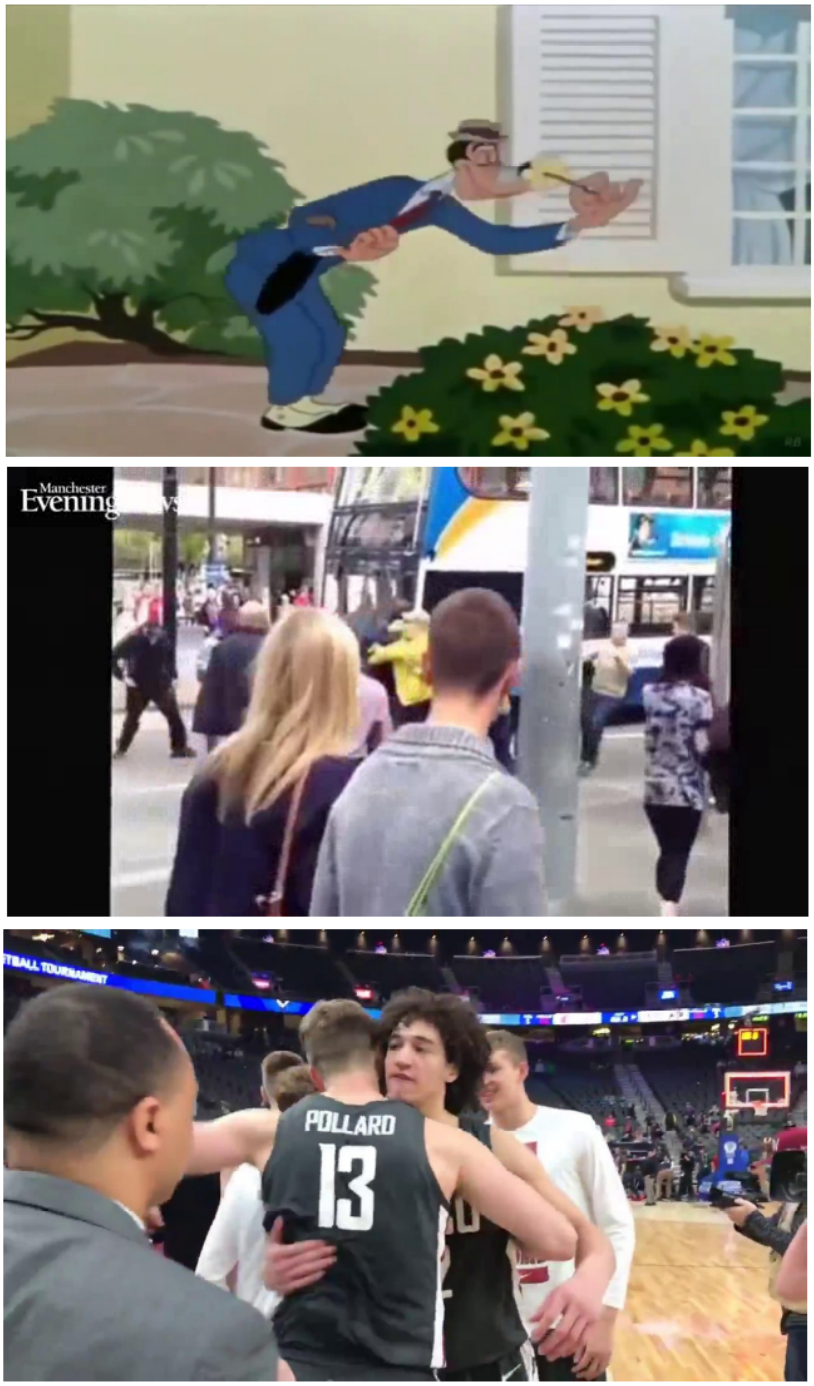

Figure 4.3 : Some non-fight samples from the dataset with different difficulty levels as (top) easy samples totally unrelated to the real-life fight actions, (middle) normal samples from real life scenarios, (bottom) hard samples depicting similar actions to fighting.

As the proposed SMFI dataset is an *in-the-wild* dataset which comprises recordings of fight moments in real-world and the recognition domain is social media content, the non-fight samples of the dataset were also collected from Twitter. There are different types of non-fight instances in the dataset as easy, normal, and hard samples. Some example images are shown in Figure 4.3. Easy samples are 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 moving in a slightly similar manner to the fight action. The sports videos with players running or throwing a ball, dancing people, some other crowded scenes, and blurry images can be categorized under hard samples. These hard samples are mostly the samples that were misinterpreted by the initial model. As many as possible hard samples were included in the dataset so that the classification will be solely based on the action displayed on the images instead of any other characteristics, such as the number of people in the scene or motion blur. The number of samples in the SMFI dataset is given in Table 4.1.

Table 4.1 : Number of fight and non-fight samples across different sources in the SMFI dataset.

	Twitter	Google	NTU CCTV-Fights [3]	Total
Fight	2247	162	330	2739
Non-fight	2642	146	164	2952
Total	4889	308	494	5691

5. EXPERIMENTS AND RESULTS

For the experimentation part, four main research questions were investigated:

- How well can one perform fight recognition on still images in the wild?
- As the available data on social media changes over time, how does this affect the performance of the trained model?
- Can a model trained on still images be used for fight recognition on videos?
- How well do the trained models generalize across different fight datasets?

The experimentation and results regarding these research questions are given in the following sections.

5.1 Fight Recognition on Social Media Images

For the fight recognition on social media images, the proposed SMFI dataset was used for the experiments. In order to compare and assess the performances of deep learning architectures on this task, several learning models were fine-tuned on the dataset. The chosen model architectures are VGG-16 [39], ResNet-50 [40], ResNeXt-50 [43], and ViT [1] as these networks are well-accepted approaches for image classification task and cover the essential concepts of it. For the performance evaluation, 10-fold cross-validation was used instead of defining a fixed test set. Considering that the samples in the dataset might be removed in time, reporting results on a fixed test set which is subject to change in the future would hamper the reproducibility of experiments. Therefore, using 10-fold cross-validation, one can achieve a more fair evaluation of techniques.

The experimental setups for each of the trained models are as follows. All networks were trained for 20 epochs.

VGG-16 [39]: Cross-entropy loss with Adam optimizer [44] was used. Weight decay was 1e-3 and the learning rate was 5e-4 for the pre-trained layers and 1e-3 for the final classification layer. The first three layers were frozen.

ResNet-50 [40] and ResNeXt-50 [43]: 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 final classification layer. First five layers were frozen.

ViT-Large-16 [1]: Cross-entropy loss with SGD (Stochastic Gradient Descent) optimizer was used. Weight decay was 1e-2 and learning rate was 3e-3.

The fight recognition results on SMFI dataset are reported in Table 5.1 as mean accuracy for 10-fold cross-validation evaluation. The results demonstrate the superiority of Vision Transformer over other networks for the fight recognition task. It outperforms the other three networks by a margin in terms of validation accuracy. Besides, the generalization ability is another significant metric for the assessment of the performance of deep learning models. The generalization of the models can be observed by comparing the train accuracies and validation accuracies. If the training accuracy is much higher than the validation accuracy, this means that the model memorizes the training data but cannot transmit the learning to unseen data. This is an unwanted behavior since the model needs to learn the overall data representation instead of only learning the training data distribution. When the gaps between the training accuracy and the validation accuracy are considered, the difference for Vision Transformer is relatively small indicating that using Vision Transformer prevents overfitting to some degree, providing a better generalization. However, the other three networks are highly overfitted meaning that they memorize the training data and fail to generalize.

Table 5.1 : Averaged 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.

Network	Train Accuracy	Test Accuracy
VGG-16	96.3%	83.0%
ResNet50	100%	87.7%
ResNext50	100%	88.3%
ViT Large 16	**96.3%**	**95.5%**

Figure 5.1 : Visual attention maps of some fight samples from the SMFI dataset. Maps were extracted from ViT and the model can successfully highlight the salient areas.

As a qualitative analysis, visual attention maps of the Vision Transformer model were extracted. The visual attention maps represent the activated regions of the feature map prior to the classification. This means that the model determines the class of the sample by looking at the highlighted areas. The activation maps given in Figure 5.1 illustrate that the Vision Transformer attends to the reasonable regions on the image to classify the samples (i.e. humans in the scene). Thus, the fight recognition is done based on the human positions instead of other characteristics of the image such as background or motion blur.

5.2 Effect of Varying Dataset Size on the Model

The dataset was collected from social media uploads as it is described in Chapter 4 and the links to the social media content were shared instead of the images themselves due to copyright issues. However, the uploaded content on social media could be removed in time by authorities or the users themselves. Especially for the violent content case, the removal speed is relatively higher since it is considered as a sensitive subject. Eventually, it might not be possible to retrieve the entire set of samples as time passes and the number of samples that can be accessed through the shared links is likely to decrease. Considering this issue, it was investigated how the trained model is affected by the removal of data.

In order to simulate the data removal pattern in various timestamps, the model was fine-tuned on different partitions of the dataset. To that end, the dataset was split into training, validation, and test sets as 70%, 20%, and 10% of whole samples, respectively. Then, five partitions were sampled as 40%, 50%, 60%, 80%, and 100% among all samples in the training and validation sets. The partitions were randomly sampled so that a consistent distribution among the partitions was provided. The test set was kept constant across the experiments for the sake of a fair comparison of the models.

Since the Vision Transformer gave the best results in Section 5.1, it was used for measuring the effect of the removal of data in time. The accuracies obtained by the network on different partitions of the data are given in Table 5.2. The evaluation results

Table 5.2 : Performance of ViT with respect to use of different amount of development 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%

indicate that the trained Vision Transformer model is robust to variations in dataset size.

When all of the data is available for training, the network achieves 95% accuracy on the test set. The training accuracy for the same setup is also 95% meaning that the model can successfully generalize. As the dataset size is decreased, the gap between the training accuracy and test accuracy increases rapidly. This indicates that as more data is available for training, the model better generalizes, and decreasing dataset size results in overfitting.

For the test accuracy, it slightly decreases as the data samples are removed but even if 60% of the samples are lost due to the deleted content on social media platforms, the accuracy was still on par with the ideal case, i.e all data samples are available. Thus, it can be claimed that the Vision Transformer can perform well on fight recognition task even if the entire dataset could not have been retrieved.

5.3 Single-frame Fight Recognition on Video-based Fight Datasets

The previous works mostly tackled the fight recognition task on video datasets using temporal information. In this scope, several fight video datasets were proposed such as Hockey [4], Movie [4], Crowd Violence [5], and Surveillance Fight [7] datasets. In order to investigate if the temporal information is necessary for the classification of these datasets, a still-image-based comparative experiment has been held on four video fight datasets:

- **Hockey Fight Dataset [4]** contains 1000 videos in total as 500 of them are fight and 500 of them are non-fight samples. The videos are 1-2 seconds long cuts from hockey game recordings where players are fighting or just playing the game.
- **Movie Dataset [4]** contains 200 videos in total where 100 of them are fight samples and 100 of them are non-fight samples. Fight samples were collected from sports games (e.g., soccer, boxing, etc.) and some Hollywood movies. Non-fight samples include casual events, such as walking and waving hands. Videos are 1-2 seconds long.
- **Crowd Violence Dataset [5]** is more focused on the crowd violence where a high number of performers are involved in the violent act. The dataset consists of 246 samples as 123 of them are fight and 123 of them are non-fight samples. Dataset was collected from real-world scenarios, such as violent actions in football games and group fighting on street. The duration of the videos varies between 1-6 seconds.
- **Surveillance Fight Dataset [7]** includes surveillance camera recordings of the fight and non-fight occasions collected from YouTube. 300 samples are included in the dataset where 150 of them are fight and 150 of them are non-fight videos. Video sequences are 1-3 seconds long.

For the still-image-based classification experiment, a single frame from each video sample was used for the classification of video datasets. The single frame was randomly sampled among the frames of videos in the respective dataset and labeled with the same label of the video. Two image classification networks were fine-tuned on the single frame data as ResNet-50 [40] and ViT [1]. The implementation details of the networks are as follows.

ResNet-50 [39]: ImageNet pre-trained network was used where the first five layers were frozen. Learning rate and weight decay were set to 1e-3 and cross-entropy loss with Adam optimizer was used.

ViT-Large-16 [1]: Large Vision Transformer with 16×16 patch size. Cross-entropy loss with SGD optimizer was used. Weight decay was 1e-2 and learning rate was 3e-3.

Table 5.3 : Results on video fight datasets. Bold ones were obtained without using temporal information. One frame was chosen randomly from each sample and these frames were classified using respective image classification networks.

	Hockey	Movie	Crowd	Surveillance
Spatiotemporal Encoder [14]	96.5%	100%	92.1%	-
ConvLSTM [13]	97.1%	100%	94.5%	-
Flow Gated Network [8]	98%	100%	88.8%	-
FightNet [45]	97%	100%	-	-
3D CNN [10]	96%	99.9%	98%	-
CNN + Bi-LSTM + attention [7]	98%	100%	-	72%
Kang et al. [46]	99.6%	100%	98%	92%
ViT (frame-based)	**98%**	**100%**	**98%**	**84.6%**
ResNet50 (frame-based)	**99%**	**99.5%**	**97%**	**76.6%**

In order to evaluate the models, 5-fold cross-validation was used as this is the common approach for fight recognition in videos literature. The accuracies of the networks fine-tuned on single frames were compared with approaches that use temporal information for video classification. Results are shown in Table 5.3.

The results reported in Table 5.3 show that it is possible to classify the video-based fight datasets using only a single frame. The single frame-based solutions with ResNet-50 [40] and Vision Transformer [1] give comparable accuracies to those approaches using temporal information across the frames. Even in some cases, it was seen that the single frame-based setup is superior to the video-based setups.

One of the reasons for the obtained results is seen as the high inter-class distribution difference between the fight and non-fight classes for Hockey and Movie datasets. Randomly chosen frames from the samples in Hockey and Movie datasets are given in Figure 5.2 and Figure 5.3, respectively. The frame distribution for the Hockey dataset demonstrates that the fight scenes in hockey games are mostly recorded from close angles since the fight actions are considered as an anomaly in the regular flow of the game and found to be intriguing for the viewers. However, for the non-fight cases, players play the game as usual and the camera records the game from a wide angle. This pattern could be a significant cue for the discrimination of fight and non-fight samples using a single frame. Similarly, for the Movie dataset, the non-fight samples are collected in controlled environments with similar backgrounds. Thus, there is a

distinct difference between the color scales and this is seen as the reason why it was possible to classify this dataset successfully using only one frame from each video. The characteristics of these two datasets explain how it is possible to achieve nearly perfect accuracies without using temporal information at all. The results also imply that there might be a bias problem for these three video-based fight datasets, Hockey, Movie, and Crowd Violence.

Differently from the Hockey, Movie, and Crowd Violence datasets, the single-frame-based classification on the Surveillance Fight dataset could not perform comparably to all video-based classification approaches. Even though the single-frame-based method outperforms the approach using CNN + Bi-LSTM + attention model [7], it falls behind the proposed approach at [46]. The work of [46] proposes a solution for video fight detection using two modules for spatial and temporal attention, which highlight the informative regions in both dimensions. It achieves better results than the single frame case making use of the temporal information, indicating that the Surveillance Fight dataset is a relatively harder dataset than the others. Using temporal information is still necessary to achieve higher performance on this dataset where spatial information is not enough on its own.

5.4 Cross-dataset Experiments

Upon the results obtained and reported in Section 5.3, a dataset bias analysis for fight recognition datasets was regarded as insightful experimentation. In this scope, the generalization of the fine-tuned models on fight recognition datasets was investigated through cross-dataset experiments. The proposed SMFI dataset was compared with other video-based fight recognition datasets, i.e Hockey [4], Movie [4], Crowd Violence [5], and Surveillance Fight [7], through the cross-dataset experiments.

For the video-based datasets, the same setup was used with the single frame experiment in Section 5.3 where a single random frame from each video is used for the training of the network. The corresponding networks for each dataset were chosen based on their classification performance on that dataset. In the end, fine-tuned ResNet-50 [40] model was used for Hockey, Movie, and Crowd Violence dataset experiments and fine-tuned Vision Transformer [1] model was used for Surveillance Fight and SMFI

Figure 5.2 : Fight class (top) and non-fight class (bottom) frames from Hockey Fight Dataset.

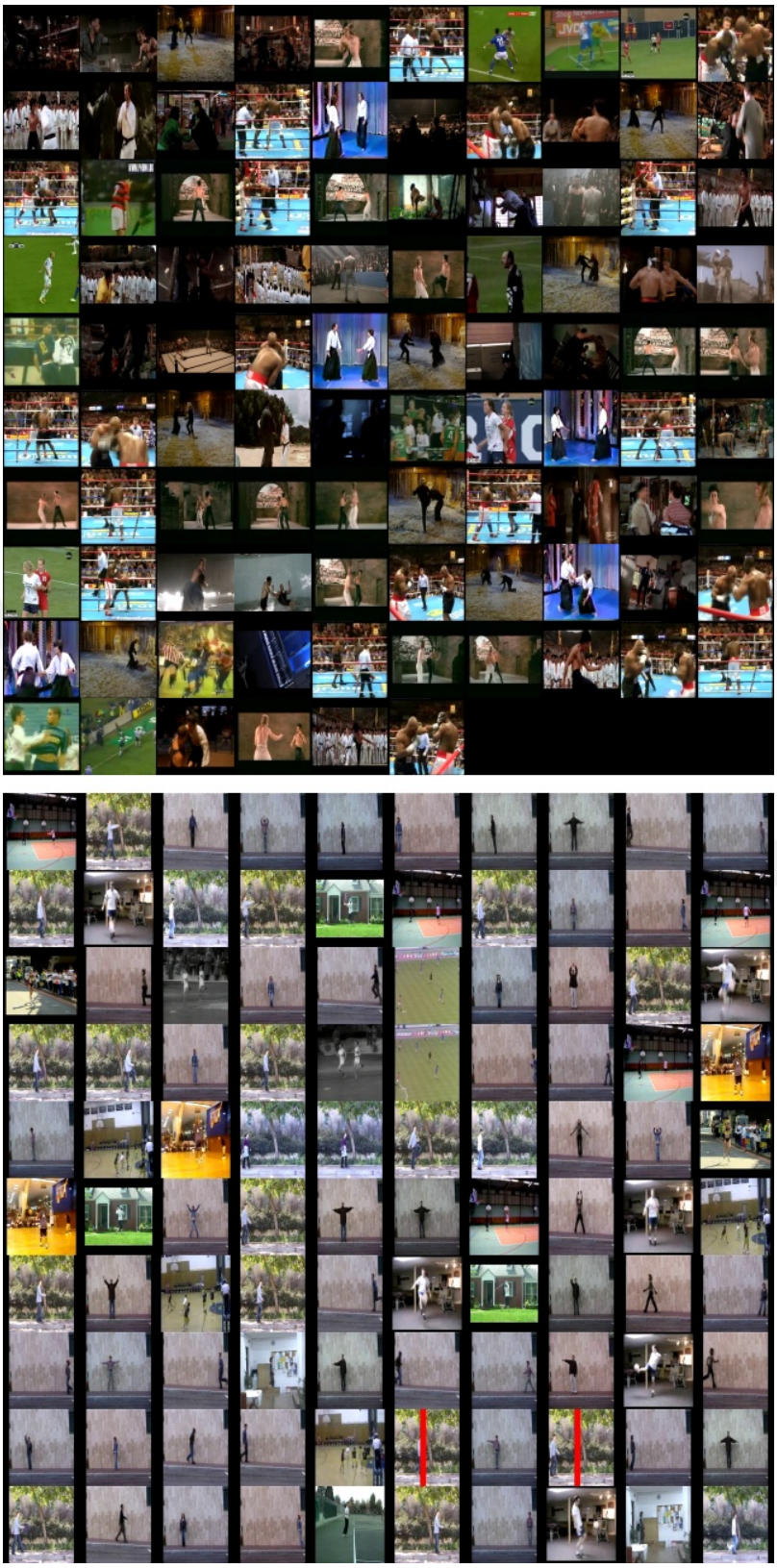

Figure 5.3 : Fight class (top) and non-fight class (bottom) frames from Movie Fight Dataset.

dataset experiments. During the testing, again single frames from each video were used for video datasets. For the SMFI dataset, the models were tested on all available samples.

Table 5.4 : Cross-dataset experiment results on video-based fight recognition datasets and proposed SMFI dataset. Rows indicate the training dataset and columns indicate the testing dataset.

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

Results for the cross-dataset experiments are reported in Table 5.4. The average accuracies indicate that the generalization ability of the Vision Transformer model fine-tuned on the SMFI dataset is higher than the three frame-based models fine-tuned on the video datasets Hockey, Movie, and Crowd Violence. This points out that the proposed SMFI dataset covers an extensive representation of the fight scenes in the real-world and the model trained on this dataset can also classify the other datasets with satisfying accuracy. Relatively lower average accuracies for video-based datasets might mean that the frames extracted from these datasets (Hockey, Movie, Crowd Violence) fail to generalize the fight recognition problem and models cannot be adapted to other domains. This once more highlights the bias problem in the available fight recognition datasets. It is also worth noting that the cross-dataset average accuracy of the Surveillance Fight dataset surpasses the SMFI. This result can be explained by the fact that the Surveillance Fight dataset is also an in-the-wild dataset and has a relatively lower bias.

6. CONCLUSION

As it is discussed in previous sections, violent content on social media can be significantly harmful to society. Thus, recognition systems for these kinds of media content were developed and used for censoring. Fight scenes are one of the most common types of violent content and are frequently shared on social media by users who witness these kinds of abnormal activities. It is also regarded as a sensitive context for people, especially for the younger generation. There were several fight recognition systems developed for recognizing fight actions from videos, which depend on temporal information available in video sequences. However, there exists a non-negligible amount of still image data on social media depicting fight scenes and no prior work has focused on the discrimination of such data. In this thesis, the task of fight recognition from still images is explored and a fight recognition system for social media case was proposed along with introducing a novel dataset that is collected from real-world social media fight images.

The introduced dataset named Social Media Fight Images (SMFI) consists of images collected from Twitter and Google and also some video frames from NTU CCTV-Fights [3]. Dataset images were collected from real-world fight scene images shared on social media platforms and various keywords in different languages are used. This dataset collection process resulted in creating an in-the-wild dataset that is not biased towards any race or geographical region and presents an extensive representation of fight scenes. The dataset was made publicly available by sharing the links to the social media contents included in the dataset.

The fight recognition from still images task was handled as a binary classification problem where the aim is discriminating between the fight and non-fight scenes. Convolutional Neural Networks and Vision Transformer were utilized for the classification and their performances were compared in terms of accuracy and generalization. In the end, it was concluded that Vision Transformer is more successful

in the classification of the SMFI dataset and generalizes the fight recognition task better than the other utilized networks. The self-attention mechanism and the robustness to the blurry images were seen as the key points to the success of the Vision Transformer on this task.

One of the problems regarding the dataset was considered as the varying dataset size in time due to the deleted content on social media platforms. Since the dataset was shared through links to the images, it might not be possible to retrieve the whole dataset in the future. Thus, the effect of varying dataset size was also investigated through data removal simulation and Vision Transformer managed to give a satisfying performance in terms of accuracy even the 60% of the data is assumed to be removed.

As the image classification networks performed with promising accuracies on fight recognition from still images task, the necessity of the temporal information for the classification of video-based fight datasets was analyzed through the single-frame based experiments. The evaluation of the networks trained on single frames from each video in video-based fight datasets indicates that the temporal information might not be a significant modality for the classification of these specific datasets, Hockey, Movie and Crowd Violence. Most of the available video datasets could be classified with nearly perfect accuracy, only using a randomly chosen single frame. In order to get an insight into the obtained results, the randomly chosen frames are examined. It is seen that the visual difference between the classes is really high and they can be easily discriminated without temporal information.

Upon the results obtained in single-frame experiments, a bias analysis on the video-based datasets is considered intriguing and cross-dataset experiments were conducted. The trained models on the video-based fight datasets perform poorly when they are tested on the other datasets. These results point out that most of the video-based datasets have a bias problem where they fail to generalize the fight recognition problem. Given that the model trained on the proposed SMFI dataset shows better performance on other datasets on average, it can be concluded that the SMFI dataset covers the domain of fight recognition in real-world scenarios better than most of the video-based fight datasets.

In conclusion, this thesis and the proposed Social Media Fight Images (SMFI) dataset set a baseline for future works in fight recognition from still images. Besides, the extensive experimental analysis reported in the thesis shows that Vision Transformer is a promising approach for fight recognition task and it exposes some significant problems regarding video-based fight recognition datasets while highlighting the generalization ability of the proposed SMFI dataset.

REFERENCES

[1] **Dosovitskiy, A.**, **Beyer, L.**, **Kolesnikov, A.**, **Weissenborn, D.**, **Zhai, X.**, **Unterthiner, T.**, **Dehghani, M.**, **Minderer, M.**, **Heigold, G.**, **Gelly, S.** *et al.* (2020). An image is worth 16x16 words: Transformers for image recognition at scale, *arXiv preprint arXiv:2010.11929*.

[2] **Imran, M.**, **Castillo, C.**, **Lucas, J.**, **Meier, P. and Vieweg, S.** (2014). AIDR: Artificial intelligence for disaster response, *Proceedings of the 23rd international conference on world wide web*, pp.159–162.

[3] **Perez, M.**, **Kot, A.C. and Rocha, A.** (2019). Detection of Real-world Fights in Surveillance Videos, *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp.2662–2666.

[4] **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*, Springer, pp.332–339.

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

[6] **Sultani, W.**, **Chen, C. and Shah, M.** (2018). Real-world anomaly detection in surveillance videos, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp.6479–6488.

[7] **Aktı, Ş.**, **Tataroğlu, G.A. and Ekenel, H.K.** (2019). Vision-based Fight Detection from Surveillance Cameras, *2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pp.1–6.

[8] **Cheng, M.**, **Cai, K. and Li, M.** (2021). RWF-2000: An open large scale video database for violence detection, *2020 25th International Conference on Pattern Recognition (ICPR)*, IEEE, pp.4183–4190.

[9] **Ding, C.**, **Fan, S.**, **Zhu, M.**, **Feng, W. and Jia, B.** (2014). Violence detection in video by using 3D convolutional neural networks, *International Symposium on Visual Computing*, Springer, pp.551–558.

[10] **Ullah, F.U.M.**, **Ullah, A.**, **Muhammad, K.**, **Haq, I.U. and Baik, S.W.** (2019). Violence detection using spatiotemporal features with 3D convolutional neural network, *Sensors*, *19*(11), 2472.

[11] **Li, J.**, **Jiang, X.**, **Sun, T. and Xu, K.** (2019). Efficient violence detection using 3D convolutional neural networks, *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, IEEE, pp.1–8.

[12] **Accattoli, S.**, **Sernani, P.**, **Falcionelli, N.**, **Mekuria, D.N. and Dragoni, A.F.** (2020). Violence detection in videos by combining 3D convolutional neural networks and support vector machines, *Applied Artificial Intelligence*, *34*(4), 329–344.

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

[14] **Hanson, A.**, **Pnvr, K.**, **Krishnagopal, S. and Davis, L.** (2018). Bidirectional convolutional LSTM for the detection of violence in videos, *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*.

[15] **Fenil, E.**, **Manogaran, G.**, **Vivekananda, G.**, **Thanjaivadivel, T.**, **Jeeva, S.**, **Ahilan, A.** *et al.* (2019). Real time violence detection framework for football stadium comprising of big data analysis and deep learning through bidirectional LSTM, *Computer Networks*, *151*, 191–200.

[16] **Ullah, W.**, **Ullah, A.**, **Haq, I.U.**, **Muhammad, K.**, **Sajjad, M. and Baik, S.W.** (2021). CNN features with bi-directional LSTM for real-time anomaly detection in surveillance networks, *Multimedia Tools and Applications*, *80*(11), 16979–16995.

[17] **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.

[18] **Zhou, P.**, **Ding, Q.**, **Luo, H. and Hou, X.** (2018). Violence detection in surveillance video using low-level features, *PLoS One*, *13*(10), e0203668.

[19] **Wu, P.**, **Liu, J.**, **Shi, Y.**, **Sun, Y.**, **Shao, F.**, **Wu, Z. and Yang, Z.** (2020). Not only look, but also listen: Learning multimodal violence detection under weak supervision, *European Conference on Computer Vision*, Springer, pp.322–339.

[20] **Pujol, F.A.**, **Mora, H. and Pertegal, M.L.** (2020). A soft computing approach to violence detection in social media for smart cities, *Soft Computing*, *24*(15), 11007–11017.

[21] **Blandfort, P.**, **Patton, D.U.**, **Frey, W.R.**, **Karaman, S.**, **Bhargava, S.**, **Lee, F.T.**, **Varia, S.**, **Kedzie, C.**, **Gaskell, M.B.**, **Schifanella, R.** *et al.* (2019). Multimodal social media analysis for gang violence prevention, *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pp.114–124.

[22] **Wang, D.**, **Zhang, Z.**, **Wang, W.**, **Wang, L. and Tan, T.** (2012). Baseline results for violence detection in still images, *2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance*, IEEE, pp.54–57.

[23] **Ikizler, N.**, **Cinbis, R.G.**, **Pehlivan, S. and Duygulu, P.** (2008). Recognizing actions from still images, *2008 19th International Conference on Pattern Recognition*, pp.1–4.

[24] **Maji, S.**, **Bourdev, L. and Malik, J.** (2011). Action recognition from a distributed representation of pose and appearance, *CVPR 2011*, IEEE, pp.3177–3184.

[25] **Qi, T.**, **Xu, Y.**, **Quan, Y.**, **Wang, Y. and Ling, H.** (2017). Image-based action recognition using hint-enhanced deep neural networks, *Neurocomputing*, *267*, 475–488.

[26] **Khan, F.S.**, **Anwer, R.M.**, **Van De Weijer, J.**, **Bagdanov, A.D.**, **Lopez, A.M. and Felsberg, M.** (2013). Coloring action recognition in still images, *International Journal of Computer Vision*, *105*(3), 205–221.

[27] **Delaitre, V.**, **Sivic, J. and Laptev, I.** (2011). Learning person-object interactions for action recognition in still images, *NIPS 2011: Twenty-Fifth Annual Conference on Neural Information Processing Systems*.

[28] **Gkioxari, G.**, **Girshick, R. and Malik, J.** (2015). Contextual action recognition with R*CNN, *Proceedings of the IEEE International Conference on Computer Vision*, pp.1080–1088.

[29] **Sharma, G.**, **Jurie, F. and Schmid, C.** (2016). Expanded parts model for semantic description of humans in still images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(1), 87–101.

[30] **Zhang, Y.**, **Cheng, L.**, **Wu, J.**, **Cai, J.**, **Do, M.N. and Lu, J.** (2016). Action recognition in still images with minimum annotation efforts, *IEEE Transactions on Image Processing*, *25*(11), 5479–5490.

[31] **Liu, L.**, **Tan, R.T. and You, S.** (2018). Loss guided activation for action recognition in still images, *Asian Conference on Computer Vision*, Springer, pp.152–167.

[32] **Yan, S.**, **Smith, J.S. and Zhang, B.** (2017). Action recognition from still images based on deep VLAD spatial pyramids, *Signal Processing: Image Communication*, *54*, 118–129.

[33] **Yu, X.**, **Zhang, Z.**, **Wu, L.**, **Pang, W.**, **Chen, H.**, **Yu, Z. and Li, B.** (2020). Deep ensemble learning for human action recognition in still images, *Complexity*, *2020*.

[34] **Ashrafi, S.S.**, **Shokouhi, S.B. and Ayatollahi, A.** (2021). Action recognition in still images using a multi-attention guided network with weakly supervised saliency detection, *Multimedia Tools and Applications*, 1–27.

[35] **Wu, W. and Yu, J.** (2021). An Improved Deep Relation Network for Action Recognition in Still Images, *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp.2450–2454.

[36] **Ma, S.**, **Bargal, S.A.**, **Zhang, J.**, **Sigal, L. and Sclaroff, S.** (2017). Do less and achieve more: Training CNNs for action recognition utilizing action images from the web, *Pattern Recognition*, *68*, 334–345.

[37] **LeCun, Y.**, **Bottou, L.**, **Bengio, Y. and Haffner, P.** (1998). Gradient-based learning applied to document recognition, *Proceedings of the IEEE*, *86*(11), 2278–2324.

[38] **Krizhevsky, A.**, **Sutskever, I. and Hinton, G.E.** (2012). Imagenet classification with deep convolutional neural networks, *Advances in neural information processing systems*, *25*.

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

[40] **He, K.**, **Zhang, X.**, **Ren, S. and Sun, J.** (2016). Deep residual learning for image recognition, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.770–778.

[41] **Vaswani, A.**, **Shazeer, N.**, **Parmar, N.**, **Uszkoreit, J.**, **Jones, L.**, **Gomez, A.N.**, **Kaiser, Ł. and Polosukhin, I.** (2017). Attention is all you need, *Advances in neural information processing systems*, *30*.

[42] **Paul, S. and Chen, P.Y.** (2021). Vision transformers are robust learners, *arXiv preprint arXiv:2105.07581*.

[43] **Xie, S.**, **Girshick, R.**, **Dollár, P.**, **Tu, Z. and He, K.** (2017). Aggregated residual transformations for deep neural networks, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.1492–1500.

[44] **Kingma, D.P. and Ba, J.** (2014). Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980*.

[45] **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*(1), 012044.

[46] **Kang, M.s.**, **Park, R.H. and Park, H.M.** (2021). Efficient Spatio-Temporal Modeling Methods for Real-Time Violence Recognition, *IEEE Access*.

CURRICULUM VITAE

Name SURNAME: Şeymanur Aktı

EDUCATION:

- **B.Sc.:** 2019, Istanbul Technical University, Computer and Informatics Faculty, Computer Engineering
- **M.Sc.:** 2022, Istanbul Technical University, Graduate School, Computer Engineering

PROFESSIONAL EXPERIENCE AND REWARDS:

- 2019-2022 Istanbul Technical University Research and Teaching Assistant
- 2019 Completed BSc at Istanbul Technical University Computer Engineering Department with high honor ranking 3rd in department.

PUBLICATIONS, PRESENTATIONS AND PATENTS ON THE THESIS:

- Akti Ş., Ofli F., Imran M., Ekenel H. K. (2022). Fight Detection from Still Images In the Wild. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision.

OTHER PUBLICATIONS, PRESENTATIONS AND PATENTS:

- Aktı Ş., Qaraqe M., Ekenel H. K. (2022). A Mobile Food Recognition System for Dietary Assessment. International Conference on Image Analysis and Processing.
- Aktı Ş., Kantarcı A., Erakın M. E., Cinik F., Timirci T., Ekenel H. K. (2022). Automatic Component Counting on PID Drawings. Signal Processing and Communications Applications Conference (SIU).
- Aktı Ş., Kamar D., Özlü Ö. A., Soydemir I., Akcan M., Kul A., Rekik I. (2022). A comparative study of machine learning methods for predicting the evolution of brain connectivity from a baseline timepoint. Journal of neuroscience methods, 109475.
- Massoli F. V., Falchi F., Kantarci A., Akti Ş., Ekenel H. K., Amato, G. (2021). MOCCA: Multilayer One-Class Classification for Anomaly Detection. IEEE Transactions on Neural Networks and Learning Systems.

- Aktı Ş., Ekenel H. K. (2020). Pre-processing of Signature Images for Detection and Verification. Signal Processing and Communications Applications Conference (SIU).
- Aktı Ş., Tataroğlu G. A., Ekenel H. K. (2019). Vision-based fight detection from surveillance cameras. International Conference on Image Processing Theory, Tools and Applications (IPTA).