RECOGNIZING HUMAN INTERACTIONS IN STILL IMAGES

RESİMLER ÜZERİNDE İNSAN ETKİLEŞİMLERİNİN TANINMASI

GÖKHAN TANIŞIK

ASSOC. PROF. DR. NAZLI İKİZLER CİNBİŞ

Supervisor

Submitted to Graduate School of Science and Engineeringof Hacettepe University
as a Partial Fulfillment to the Requirements
for the Award of the Degree of Doctor of Philosophy
in Computer Engineering

2020

To my wife and son...

ETHICS

In this thesis study, prepared in accordance with the spelling rules of Institute of Graduate Studies in Science of Hacettepe University,

I declare that

- all the information and documents have been obtained in the base of the academic rules.
- all audio-visual and written information and results have been presented according to the rules of scientific ethics
- in case of using others works, related studies have been cited in accordance with the scientific standards
- all cited studies have been fully referenced
- I did not do any distortion in the data set
- and any part of this thesis has not been presented as another thesis study at this or any other university.

19/08/2020

GÖKHAN TANIŞIK

YAYINLANMA FİKRİ MÜLKİYET HAKLARI BEYANI

Enstitü tarafından onaylanan lisansüstü tezimin/raporumun tamamını veya herhangi bir kısmını, basılı (kağıt) ve elektronik formatta arşivleme ve aşağıda verilen koşullarla kullanıma açma iznini Hacettepe üniversitesine verdiğimi bildiririm. Bu izinle Üniversiteye verilen kullanım hakları dışındaki tüm fikri mülkiyet haklarım bende kalacak, tezimin tamamının ya da bir bölümünün gelecekteki çalışmalarda (makale, kitap, lisans ve patent vb.) kullanım hakları bana ait olacaktır.

Tezin kendi orijinal çalışmam olduğunu, başkalarının haklarını ihlal etmediğimi ve tezimin tek yetkili sahibi olduğumu beyan ve taahhüt ederim. Tezimde yer alan telif hakkı bulunan ve sahiplerinden yazılı izin alınarak kullanması zorunlu metinlerin yazılı izin alarak kullandığımı ve istenildiğinde suretlerini Üniversiteye teslim etmeyi taahhüt ederim.

Yükseköğretim Kurulu tarafından yayınlanan "***Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge***" kapsamında tezim aşağıda belirtilen koşullar haricince YÖK Ulusal Tez Merkezi / H. Ü. Kütüphaneleri Açık Erişim Sisteminde erişime açılır.

- □ Enstitü / Fakülte yönetim kurulu kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren 2 yıl ertelenmiştir.
- □ Enstitü / Fakülte yönetim kurulu gerekçeli kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren .... ay ertelenmiştir.
- □ Tezim ile ilgili gizlilik kararı verilmiştir.

19/08/2020

GÖKHAN TANIŞIK

ABSTRACT

RECOGNIZING HUMAN INTERACTIONS IN STILL IMAGES

Gökhan TANIŞIK

Doctor of Philosophy, **Computer Engineering**
Supervisor: Assoc. Prof. Dr. Nazlı İKİZLER CİNBİŞ
19/08/2020, 101 pages

Recognizing human interactions in still images is quite a challenging problem since compared to videos, there is only a glimpse of interaction in a single image. In this thesis, we explore the role of two components that provide descriptive features upon recognizing interactions in still images: the role of human faces and poses.

As for the role of human faces, we explore whether the facial regions and their spatial configurations contribute to the recognition of interactions. In this respect, our method involves the extraction of several visual features from the facial regions, incorporating scene characteristics and deep features to the recognition. Extracted multiple features are utilized within a discriminative learning framework for recognizing interactions between people. Our designed facial descriptors are based on the observation that relative positions, size, and locations of the faces are likely to be essential for characterizing human interactions. Since there is no available dataset in this relatively new domain, a comprehensive new dataset that includes several images of human interactions is collected. Our experimental results show that faces and scene characteristics contain vital information to recognize interactions between people.

On behalf of exploring the role of human poses upon recognizing interactions, we propose a multi-stream convolutional neural network architecture, which fuses different levels of human pose information to recognize human interactions better. In this context, several pose-based representations are explored. Experimental evaluations in an extended benchmark dataset show that the proposed multi-stream pose Convolutional Neural Network successfully discriminates a wide range of human-human interactions. Moreover, when used in conjunction with the overall context, human poses provide discriminative cues about human-human interactions.

Keywords: human-human interactions, facial features, poses, convolutional neural networks

ÖZET

RESİMLER ÜZERİNDE İNSAN ETKİLEŞİMLERİNİN TANINMASI

Gökhan TANIŞIK

Doktora, Bilgisayar Mühendisliği

Danışman: Assoc. Prof. Dr. Nazlı İKİZLER CİNBİŞ

Ekim 2018, 101 sayfa

Resimler üzerinde insan etkileşimlerini tanımak oldukça zordur. Özellikle resimler ve videolar kıyaslandığında, resimde etkileşime dair sadece küçük bir an bulunur. Bu tez kapsamında, resimler üzerinden insan etkileşimlerinin tanınması için iki bileşenin rolü incelenmiştir: yüzler ve pozlar.

Yüzlerin rolünü incelemek adına, resimlerde yüzlerin bulunduğu bölgelerin ve bu yüzlerin uzamsal yerleşimlerinin etkileşimleri tanımlamada ne kadar katkı sunduğunu inceledik. Bu doğrultuda, geliştirdiğimiz yöntem yüzlerden elde edilen bir çok niteliği, çeşitli sahne nitelikleri ve derin öğrenme niteliklerini bir araya getirir. Elde edilen bu nitelikler bir öğrenme sistemi içerisinde, birlikte işlenerek insan etkileşimlerinin ayırt edilmesini sağlar. Tanımladığımız yüz nitelikleri; yüzlerin birbirine göreceli uzaklıkları, boyutları ve konumları gibi etkileşimleri birbirinden farklı kılabilecek nitelikleri kapsar. Bu çalışma kapsamında kullanabileceğimiz önceden tanımlı bir veri kümesi bulunmadığından, çeşitli insan etkileşimlerini barındıran yeni ve zorlu bir veri kümesi oluşturulmuştur. Yapılan deneyler, yüzlerin ve sahne niteliklerinin insan etkileşimlerini tanımlamada önemli bir paya sahip olduğunu göstermektedir.

Etkileşimleri tanımlamada pozların rolünü araştırmak için de, çoklu-akışlı bir evrişimsel sinir ağları mimarisi tanımladık. Bu mimari her bir akıştan elde edilen farklı seviyelerdeki poz bilgilerini bir araya getirir. Bu bağlamda, çeşitli poz tabanlı gösterimler incelenmiştir. Genişletilmiş bir veri kümesi üzerinde yapılan deneyler, önerilen yöntemin insan etkileşimlerini sınıflandırabilmek için oldukça başarılı olduğunu göstermiştir. Bu poz tabanlı gösterimler ve orjinal resimler bir arada kullanıldığında başarımın daha da arttığı gözlenmiştir.

Anahtar Kelimeler: insan etkileşimlerini tanıma, yüz nitelikleri, pozlar, evrişimsel sinir ağları

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my supervisor Assoc. Prof. Dr. Nazlı İKİZLER CİNBİŞ for her valuable advice and guidance. At every stage of this thesis, she supported me with her knowledge, experience, motivation, and encouragement.

I thank the members of my thesis progress committee, Prof. Dr. A. Aydın ALATAN and Assoc. Prof. Dr. İ. Aykut ERDEM for their guidance, leading comments and positive attitude. Moreover, I would like to thank to my thesis committee members Prof. Dr. Pınar DUYGULU ŞAHİN, and Assoc. Prof. Dr. Sinan KALKAN for reviewing this thesis and giving insightful comments.

I would also thank to members of Hacettepe University Computer Vision Laboratory (HUCVL) for providing me a pleasant atmosphere for working. I would also like to extend my deepest gratitude to Çağdaş Baş for his endless patience and helps with using computers in the laboratory.

I would like to thank my parents Memet and Müzeyyen, for supporting and encouraging me throughout my educational life.

Finally, I am deeply grateful to my wife, Yasemin, for her best wishes and support.

This research was supported in part by TUBITAK Career Development Award 112E149.

CONTENTS

FIGURES

TABLES

1. INTRODUCTION

1.1 Motivation

Human activities are a hot topic in the computer vision community. Studies start from simple primitives such as understanding gestures and goes up to complex group activities. Among these topics, this thesis attack on the problem of recognizing human-human interactions in still images.

Human interactions have been a research area of social psychology for a long time. From the perspective of computational vision, the focus of this thesis is on the observable forms of interactions. In terms of Social Psychology, Newcomb *et al.* [17] define these interactions as; "*any set of observable behaviors on the part of two or more individuals when there is a reason to assume that in some part those persons are responding to each other. What all these observable forms of interaction have in common is a sequence of behaviors on the part of two or more persons.*" Although being a sequence of behaviors is claimed to be the common property of interaction, we believe that images are also capable of describing an interaction. However, it is a challenging task because of the following reasons: 1) Some interactions (such as dining, partying, speech, etc.) can have a vast deal of intra-class variety. 2) While some interactions (such as kissing and handshaking) are usually performed by two people, some interactions (such as dining, partying, or speech) may involve hundreds or thousands of people, making the granularity of the problem quite diverse. 3) There is no constraint on image characteristics, where the camera position, light amount, or clutter can vary dramatically from image to image. 4) There can be other people in the scene not related to the interaction, which makes it harder to focus on the actual interaction. 5) Lack of temporal movement data makes it harder to diversify the interaction from the background.

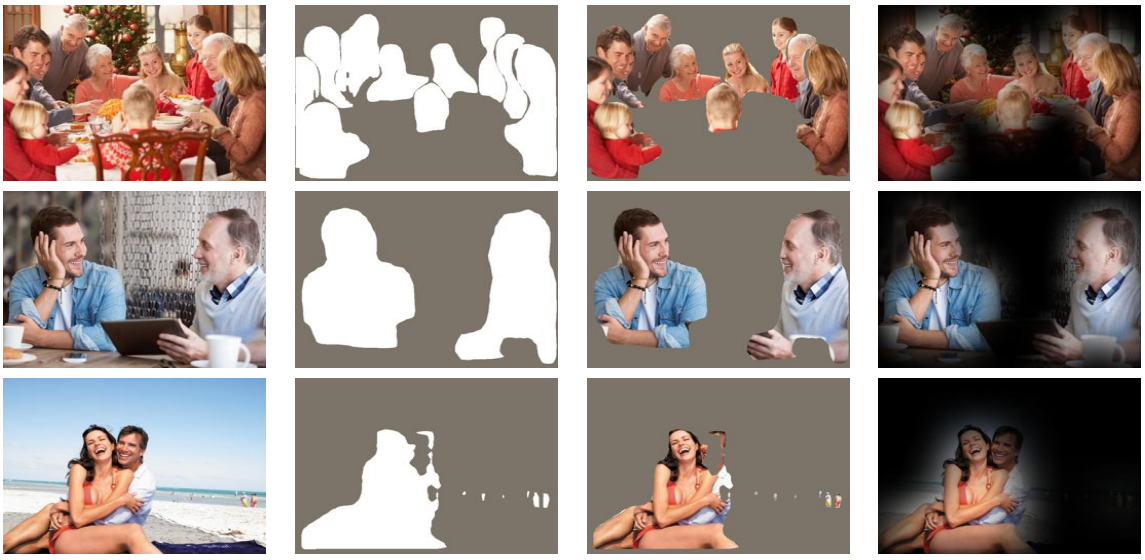
Although recognizing human actions [7, 18, 19, 20, 21] or human-object interactions [22, 23, 24, 25, 26] have attracted many researchers, recognizing human-human interactions is a lesser extent, especially when it comes to working on still images. There are a few key features to identify human-human interactions such as context, scene, and poses. Among these features, we can argue that human poses are the main ingredient of an interaction being performed since the context and scene information can change dramatically. In contrast, the poses should remain as a member of a recognizable set of discriminative elements. That being said, we also believe that faces can tell about the interaction, too. Regarding these, in this work, we address the recognition of human interactions in sill images. We propose

FIGURE 1.1: Faces in interactions. The original images are shown in the bottom row and the facial regions are shown on top. In this work, we explore whether we can predict the type of human interactions in an image based on descriptors extracted from faces, and their spatial layouts. As can be seen from this figure, without using any context or scene information, recognizing interactions by only face information can be quite a difficult task even for humans. Our results show that using facial descriptors together with global scene descriptors yield promising results for human interaction recognition in still images.

two methods to solve this problem, which investigates the role of human faces and poses in a scene.

In the first part, we study the role of human faces in human interactions. Several novel visual features are proposed that captures the intrinsic layout and orientation of face regions. Faces tend to play a great role in characterizing human interactions. People look at each other when they are talking, faces come together when people are kissing, and more. Figure 1.1 includes some examples. [27] use faces in video sequences to describe interactions in first-person (egocentric) videos. Based on inspiration from this work, we explore whether facial features can also help discriminate multi-person interactions in still images.

Another reason to explore facial features is that face detection technology is considerably advanced and can locate plenty of faces in images, especially those that are not too small or significantly occluded. Our designed descriptors are based on the observation that relative positions, size, and locations of the faces are likely to be essential for characterizing human interactions. To extract these descriptors, we first use a face detector. We also estimate the orientations of the faces, if possible, using the face detector of [2]. In this way, we estimate the size, spatial location, and the orientation of the face in a [-90°,90°] range with 15° resolution. We then make use of these features to propose image-level facial interaction descriptors. For recognition, we combine these multi-person facial descriptors with standard scene descriptors extracted globally from the images. In this context, we also investigate the effect of the state-of-the-art deep learning-based features, aka, Convolutional Neural Network (CNN) features to this new problem domain.



a b c d

FIGURE 1.2: **a** shows the original images from our dataset. **b** is Pose Mask Images, where the poses in the images are masked, and the rest of the image is set to the ImageNet mean pixel values. **c** Pose Only Images, where pose regions are left as is, and the rest of the image is set to the ImageNet mean pixel values. **d** Highlighted Poses; pose regions are highlighted with a 2D gaussian mask. Note that poses are detected automatically using MaskRCNN [1]. The figure is better viewed by zooming in.

Since there is no available dataset in this relatively new domain, a new and comprehensive dataset, which includes a total of ten human interaction classes, such as boxing, dining, kissing, partying, talking, is collected. This dataset has also been enriched with the manual annotations of the ground truth face locations and orientations to facilitate further comparisons. We name this dataset as Human Interaction Images (HII).

As for the second part, we investigate the role of human poses in interactions. A multi-stream convolutional neural network (CNN) architecture is formulated that focuses mainly on pose information. To this end, we explore several forms of pose-based inputs: 1) Pose Masks, where poses are extracted and represented as binary masks. 2) Pose Only Images: poses cropped out of original images. 3) Pose Highlighted Images: Instead of directly cropping poses, they are highlighted with a Gaussian mask where the background regions are de-emphasized. Figure 1.2 shows examples of these pose based representations. We propose to use a single stream for each of these inputs and fuse the knowledge from each stream within a multi-stream CNN architecture. Our experiments demonstrate that these pose sets are good enough to describe and discriminate interactions to a great extent. Besides, our multi-stream architecture yields better results compared to single-stream counterparts.

The benchmark datasets which focus on human interaction recognition are mainly limited

to videos [28, 29, 30, 31, 32, 33, 34, 35]. While there are several recent datasets proposed ([36, 37, 38]), there are some limitations, such as [36] is not available for download; [37] contains only three classes: family, group and wedding photos; [38] defines classes as touch codes of interactions such as *hand touch hand, hand touch shoulder, hand touch torso* which semantically differs from our problem. Our HII dataset is the only one focusing solely on human-human interactions. Therefore, to test our proposed approach, we use this benchmark dataset and also collect an extended version of it, since it is a relatively small dataset for training deep models. This new version of HII is called as HIIv2.

1.2 Contributions of the Thesis

- We explore the contribution of face features to interaction recognition in images;
  - We present novel descriptors based on facial regions for human interaction recognition.
  - Our experimental results show that deep learning-based features are effective in the recognition of human-human interactions in images.
  - Our proposed facial features that aim to encode the relative configurations of faces provide useful information, especially when combined with global image features.
  - We collect a new image dataset for human interaction categorization which includes multi-person interaction instances.
  - Experiments show that faces provide complementary information for recognition.
- We explore the contribution of poses to human-human interaction recognition in images;
  - We propose to use a variety of different pose-based inputs for interaction recognition.
  - We formulate a multi-stream pose CNN that aggregates the information from different pose streams.
  - An extended benchmark dataset for human-human interaction recognition in images is collected.
  - Experiments show that paying more attention to poses has a positive effect on recognizing interactions.

1.3 Organization of the Thesis

Thus far, a summary of the thesis is given. The rest of it is as follows:

Chapter 2 summarizes the background of the methods and techniques used in this thesis. To be more specific, some of the handcrafted image descriptors and a classification method is elaborated. Convolutional Neural Networks (CNNs), and some of the most common CNN architectures are summarized.

In chapter 3, a review of the literature on human action, human-object interaction, and human-human interaction recognition methods in still images are given.

Chapter 4 describes our facial descriptors for human interaction recognition on images. The formulation of each feature is elaborated in detail.

In Chapter 5, gives the details of our proposed pose-based inputs. Moreover, combining these inputs to form a multi-stream CNN architecture is presented.

Chapter 6 introduces the datasets collected for evaluation. Details of both HII and HIIv2 is presented. The collection and annotation process of the images are also elaborated.

Chapter 7 presents the implementation details and experiments based on the proposed methods. A comprehensive evaluation of the methods and comparison to related work is presented.

Finally, in Chapter 8, we discuss the conclusion of this work and possible feature directions.

2. BACKGROUND

This thesis is focused on recognizing human-human interactions in still images. To be more specific, it aims to classify a given image to a predefined interaction. In this context, two independent works are introduced; one with the traditional hand-crafted features and traditional classification methods, and the other with recently popular Convolutional Neural Networks. This section gives a summary of the methods and techniques used in this thesis.

2.1 Support Vector Machines (SVM)

SVM [39], with a solid theoretical background, is one the most used and successful classification algorithm among the traditional methods. It can be trained with a small number of training samples, which is effective even when the number of dimensions is higher than the number of samples. However, it suffers when the number of samples is much higher than the number of feature dimensions [40].

The main idea behind SVM is to find a hyperplane in N-dimensional space(N is the number of features), which distinctly separates the data points of different classes. With margin being the distance between the hyperplane and nearest data point from either data set, its goal is to find a hyperplane with maximum possible margin.

2.2 Face Detection

In our first work, we investigate the role of faces in human interactions. To create a fully automated approach, we have used a face detection algorithm. Face detection algorithms have reached human performance with the rise of deep learning. However, these algorithms have yet to be discovered at the time of our research. Thus, we used the state-of-the-art algorithm [2] known in this research. Therefore, face detection algorithm [2] is summarized under this section.

[2] presents a unified model for face detection, landmark, and pose estimation. It is based on a mixture of trees of parts. See Figure 2.1 for an example output. Mixture represents the pose, and parts represent the landmarks of the face. Each mixture uses a shared pool of parts V. A tree-structured model is introduced where each tree is formulated as $T_m(V_m, E_m)$. Here m indicates a mixture and $V_m \subseteq V$. For an image I, and pixel location of a part $l_i = (x_i, y_i)$,

FIGURE 2.1: Sample output from [2] with unified face detection, landmark estimation, and pose estimation results.

it tries to maximize the score of part configurations $L = l_i : i \in V$ as:

$$S(I, L, m) = App_m(I, L) + Shape_m(L) + \alpha^m \tag{1}$$

$$App_m(I, L) = \sum_{i \in V_m} w_i^m \cdot \phi(L, l_i) \tag{2}$$

$$Shape_m(L) = \sum_{ij\ in E_m} a_{ij}^m dx^2 + b_{ij}^m dx + c_{ij}^m dy^2 + d_{ij}^m dy \tag{3}$$

Equation 2 sums the appearance evidence for placing a template w_i^m for part i, tuned for mixture m, at location l_i. $\phi(I, l_i)$ on the other hand is the feature vector extracted from the location l_i with a feature encoding method such as Histogram of Gradients (HoG) [41]. Equation 3 calculates a score for the spatial positioning of parts L, where $dx = x_i - x_j$ and $dy = y_i - y_j$ are distance between parts i and j. Each term in the sum can be considered as a spring that suggests spatial constraints between a part pairs. The parameters (a, b, c, d) correspond to each spring and rest location's rigidity.

FIGURE 2.2: Mixture-of-trees model from [2] to encode topological changes due to viewpoint. Red lines stand for the springs between landmark pairs.

The inference of the model corresponds to maximizing $S(I, L, m)$ in Equation 1 over L and m:

$$S^*(I) = \max_x [\max_L S(I, L, m)]$$

To summarize, for each mixture, look for the best configuration of parts.

2.3 Scene Features

A classification method needs discriminative features to separate instances that belong to different sets. Depending on the problem, features may belong to local regions or the whole image in the image domain. In the whole image case, the features are usually called global or scene features. In this section, some of the scene feature extraction methods used in this thesis are described.

2.3.1 Gist Descriptors

GIST descriptors [42] are based on a low dimensional representation of the scene, which is called Spatial Envelope [42]. A set of perceptual dimensions representing the dominant spatial structure of a scene are proposed, such as roughness, naturalness, openness, ruggedness, and expansion. It is shown that it is possible to estimate these dimensions using spectral and coarsely localized information. The image is divided into a $4x4$ grid first, and for each cell, orientation histograms are computed. To make even sized grids, it is common to first resize the image into a square shape whose width usually ranges from 32 to 128 pixels. Since the descriptor represents only the general details of the image, small sizes are sufficient. The feature histograms are computed at four different scales and eight orientations, making a $4x8 = 32$ feature size for each grid and $32x16 = 512$ for the whole image.

2.3.2 Scale Invariant Feature Transform (SIFT)

Scale Invariant Feature Transform is produced by Lowe [43] that is used to detect distinctive local image keypoint descriptors. The descriptors are translation, rotation, and scale-invariant, which creates a desirable representation of the features. In general, it has two main steps: detecting the interest points in an image and creating a feature vector from these points. To capture the main details of a given image, Gaussian blur operators are applied in different scales. The scale corresponds to σ of Gaussian kernel; the image gets more blurred while the σ increases. Then, the Difference of Gaussian (DoG) operation is applied to each consecutive scales of the image. This captures important key points, such as corners. To make sure it'is more robust, this process is kept on through multiple octaves, where an octave means half the size of the previous original image. Local minima/maxima points on the DoG images are set as interest points.

In the second part of the process, descriptors are defined based on these interest points. A $16x16$ image patch, centering the key point, is extracted, and gradients of this patch are calculated. Roughly, the most frequent gradient orientation is called as the dominant orientation. Then, the patch is divided into $4x4$ cells, and for each of these cells, a histogram of orientations for 8 orientation bins spaced apart by $45°$ is calculated. The orientations are calculated relative to the dominant orientation that makes the SIFT features orientation invariant. Histograms from each cell are concatenated to form a 128 dimensional vector called the SIFT descriptor vectors.

2.3.3 Bag Of Words (BoW)

The bag-of-words method is coming from the text-retrieval world. On a training image set, the aim is to create a codebook of the most common features extracted from each image in the set. To create a codebook, usually, a *k-means* clustering is performed on the features, where *k* defines the size of the codebook. To represent an image, a histogram of size *k* is created. Each bin of the histogram counts the number of occurrences of a feature that belongs to the cluster at that index.

In this thesis, the BoW codebook is created on the SIFT features of the image. BoW alone encodes the features without any spatial information. To overcome this problem, Spatial Pyramid Matching (SPM) [44] method introduced. Depending on the height of the pyramid, SPM splits the image into $2^h x 2^h$ grids for each height level h where $0 <= h < H$, and

H is the highest depth. Each feature belongs to the same bin – element of BoW histogram – is counted for each cell of every level, composing sub histograms for each cell. These histograms are then concatenated to create a whole representation. On the other hand, to decrease the contribution of higher levels of pyramids, a weight is given to each level. The weight gets lower while the levels go up.

2.4 Convolutional Neural Networks (CNNs)

CNN's are very powerful tools for visual recognition tasks. They became popular with the release of AlexNet architecture [3]. AlexNet achieved massive success on the ImageNet [45] dataset that took the whole world's attention.

In a nutshell, a CNN architecture is a sequence of layers. Each layer takes the input, performs a transformation, and passes it to the next layer. In the following sub-sections, a summary of main CNN Layers and some of the common CNN architectures are given.

2.4.1 Main CNN Layers

This section summarizes the main CNN layers: Convolution, Pooling, and Fully-Connected layers.

2.4.1.1 Convolutional Layers

In a fully connected network, all nodes in a layer are fully connected to all nodes in the previous layer. This is the bottleneck of Neural Networks since it requires too many operations and memory, which is not very feasible for current computers. On the other hand, for an image, a fully connected network would not be beneficial since most of the image features are local.

A Convolutional Layer applies K number of filters ($F * F$) to a given input ($H_1 * W_1 * D$) and produces an output ($H_2 * W_2 * K$). K corresponds to the depth of the filter. Each filter k_i, where $0 < k < K$, tries to learn a different aspect of the input. In addition to K, two other hyper-parameters control the output size of the layer: stride and zero-padding. Stride (S) defines the number of jumps when we move filters for the next input. Zero-padding$P$ is used to control the output size and applied as adding zeros around the input's border. Thus, the output size can be calculated as $W_2 = (W_1 - F + 2P)/S + 1$ and $H_2 = (H_1 - F + 2P)/S + 1$.

However, most common networks resize the input into a predefined square shaped size, or in other words, they use equal height and widths.

2.4.1.2 Pooling Layers

Although CNN's contains fewer parameters than fully connected neural networks, the number of parameters can increase to large amounts very quickly. One common way to reduce the parameters is to use a Pooling Layer. Another advantage of Pooling Layers is that they also help to control overfitting. A standard pooling layer operates individually on every depth slice of the input uses MAX operation to reduce the size spatially.

In general, a pooling layer has two parameters. The first one is the size of the filter (F), and the second is stride (S), which defines the spatial jumps between two consequent inputs. For an input with size $W_1 * H_1 * D_1$, the Pooling Layer with parameters F and S produces the output of size $W_2 * H_2 * D_1$ where $W_2 = (W_1 - F)/S + 1$ and $H_2 = (H_1 - F)/S + 1$. Usually, a filter with a size $2x2$ is applied with a stride 2 which removes $\%75$ of the input.

2.4.1.3 Fully-Connected Layers

A Fully-Connected Layer used at the end of a network architecture to classify the inputs. The term "Fully-Connected" refers to every neuron from the previous layer is being connected to every neuron on the next layer. It operates on the output of the preceding layer and determines which features most related to a particular class.

2.4.2 CNN Architectures

Most common CNN Architectures stack a few *conv-relu* layers, followed by *pool* layers. This pattern repeats until the image gets spatially small enough. Then, a few *fc* layers are used at the end to classify the input. Here *conv, pool, relu* and *fc* respective correspond to Convolutional Layer, Pooling Layer, Rectified Linear Unit (RELU), and Fully-Connected layer.

A RELU simply applies a thresholding operation at zero, such as *max(0,x)*, to each element of the input. It is a non-linear operation. The output size remains the same as the input size.

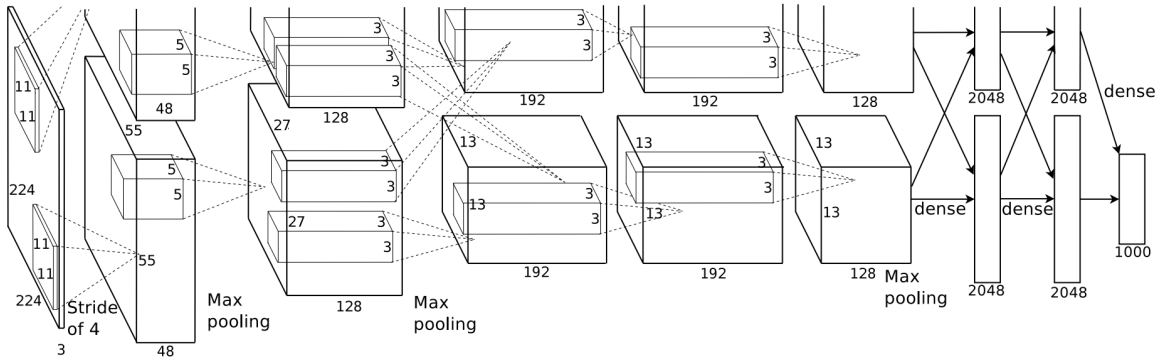


FIGURE 2.3: Original AlexNet architecture presented in [3], demonstrating the separation of the layers into two GPUs. The GPUs share information only at certain layers.

There a lot of known CNN architectures. Some of the most common ones are LeNet[46], AlexNet [3], VGG [47], GoogleNet [48], ResNet [4], and DenseNet[5]. Since AlexNet, ResNet and DenseNet are used for evaluations of this thesis, they are elaborated in the following sections.

2.4.2.1 AlexNet

AlexNet [3] is the winner of ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [45] on 2012. For the time being, the depth of the network was so deep, which made it require 2 GPUs to fit all the parameters into memory. It has eight layers in total, the first five are Convolutional, and the last three are Fully-Connected. The original architecture presented in [3] is demonstrated in figure 2.3 [1].

2.4.2.2 ResNet

ResNet [4] is one of the most successful recent CNN architectures that offer high accuracy recognition rates for the image classification problem. The idea behind its success is the residual learning architecture, which is made possible by the usage of *skip-connections*. In general, skip-connections refer to those connections skipping one or more layers. See Figure 2.4 for a visualization of skip connections. In ResNet, the authors simply perform *identity* mapping. More formally, an input x goes through *conv-relu-conv* series. This produces some $F(x)$ which is added to the original input x, resulting an $H(x) = F(x) + x$. $H(x)$ would be equal to $F(x)$ is a regular CNN without skip-connections. Instead of computing $F(x)$ as a straight mapping from x to $F(x)$, the term to be added to the original x is being computed via the layers. In other words, the residual block is computing a small change

[1] The figure 2.3 is reprinted, with permission, from [3] ©2012 NIPS.

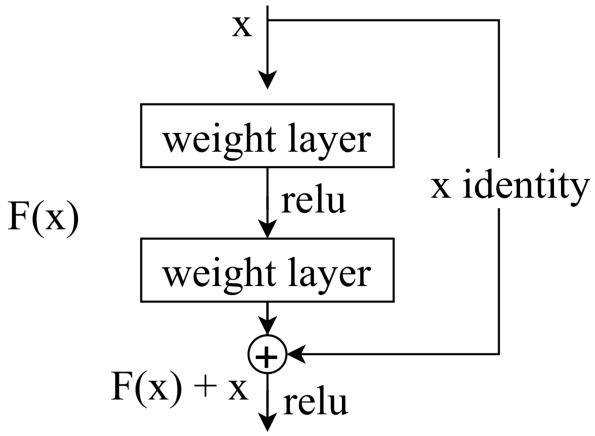


FIGURE 2.4: Skip connections of ResNet [4].

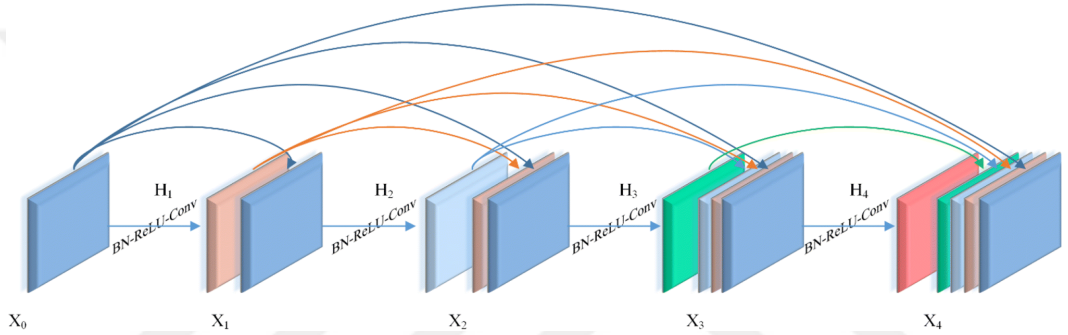


FIGURE 2.5: Sample DenseNet [5] architecture: A 5-layer dense block. Every layer takes all previous feature maps as input.

to the original input x, making the CNNs more easily trainable by bypassing the vanishing gradient problem.

2.4.2.3 DenseNet

DenseNet [5], similar to ResNet, uses the output from previous layers to ensure maximum information flow between layers in the network. The layers get directly connected to every other layer in a feed-forward fashion. For example, an architecture with L layers would have $L(L+1)/2$ direct connections. In contrast to ResNet, it has two main differences. The first one is that DenseNet uses concatenation instead of summation for fusing the data from previous layers. Another one is that a layer gets connected to every preceding layer instead of only the previous one. Because of its dense connectivity pattern, it is referred to as Dense Convolutional Network (DenseNet). See Figure 2.5 [1] for an example of dense layers.

[1]The figure 2.5 is reprinted, with permission (license number: 4847561252631), from [49] ©2020 Springer Nature.

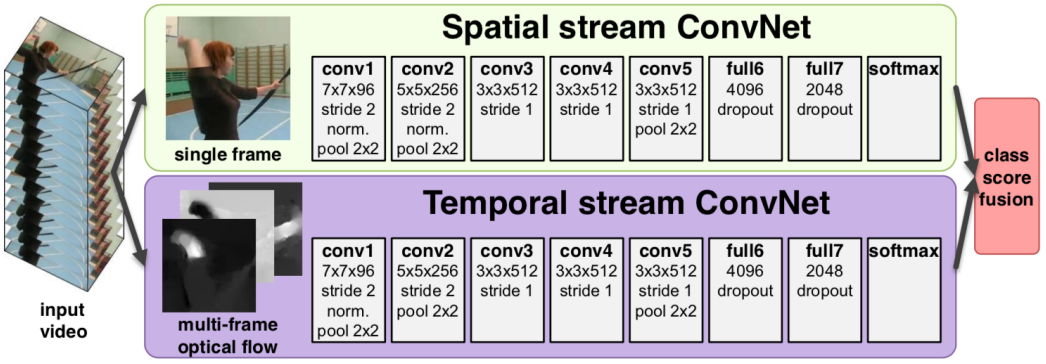


FIGURE 2.6: Two-stream architecture for action classification in videos [6].

2.4.3 Fine Tuning and Transfer Learning

A CNN architecture requires a great amount of data to learn a general representation of the input set. It gets even worse when the depth of the network increases. Data augmentation is one way to deal with, yet it only helps an extent. Another option is, instead of training from scratch, using an already trained model and train it some more. Here comes the two terms *Fine Tuning* and *Transfer Learning*. Although they are sometimes used interchangeably, *Fine Tuning* indicates taking an already learned model and training it some more on possibly different data. It is common to freeze some early layers and learn only final layers; however, it is not a necessity. *Transfer Learning*, quite similar to *Fine Tuning*, uses a part or all of the already learned model. This time, the aim is to apply the model on a different, yet somehow related task. The model is trained some more to fit on the new task.

2.4.4 Multi-Stream Neural Networks

Neural Networks usually perform great on classification and detection tasks, especially on image data. The problem arises when we try to fuse information from multiple sources or aspects. For example, it may be simple to train a model for action-recognition tasks by the spatial or temporal data. It would be quite beneficial to use all the data available to increase the performance of the network. With this respect, Simonyan *et al.* [6] proposed two-stream CNN architecture for action recognition in videos. See figure 2.6[1]

Although Simonyan's [6] work is a good example, it requires a late fusion of streams, which

[1]The figure 2.6 is reprinted, with permission, from [6] ©2014 NIPS.

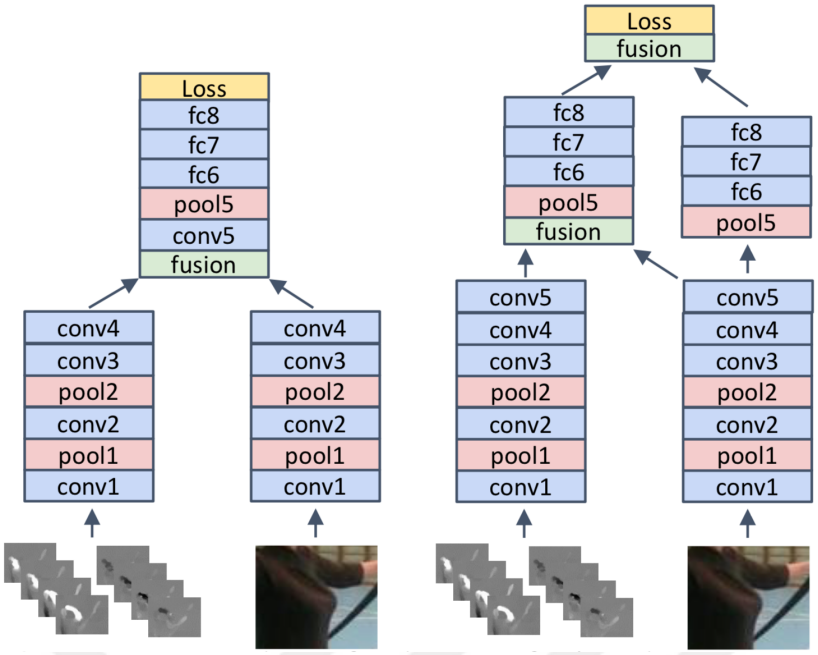


FIGURE 2.7: Two different examples to two-stream architectures which are trained end-to-end. In this respect, they are fused before the loss layer [7].

means each stream is trained separately. Thus, a better approach would be to train an end-to-end architecture that learns the model as a whole such as [7, 50]. See Figure 2.7[1] for an end-to-end trained architecture example. The fusion of the streams is done before the loss layer with a fusion layer such as Convolutional, Sum, or Max layers.

2.5 Human Detection and Segmentation

Humans are the only crucial components of the interaction. Thus it is extremely useful to detect people in the images. Detecting usually refers to find the bounding boxes surrounding each person separately. To further examine the interaction, human poses estimation or a segmentation would be very beneficial. The term *pose estimation* meant to localize human joints or, in other words, key points such as elbows, and wrists. Our work uses the segmentation masks only. Thus we omit the pose estimation part here. With this respect, we use the state-of-the-art object detection algorithm Mask R-CNN [1], which also does the instance segmentation at the same time. The following subsections summarize Mask R-CNN and Faster R-CNN [51], which is the base research of Mask R-CNN.

[1]The figure 2.7 is reprinted, with permission, from [7] ©2016 IEEE.

2.5.1 Faster R-CNN

Faster R-CNN is one of the state-of-the-art object detection algorithms that depend on region proposal to estimate object locations. It has two networks, a region proposal network (RPN) to predict possible object regions, and second to classify objects in those regions. These two networks also share the same convolutional features, which makes the method fast and powerful.

Given an image, it is provided as input to conv layers, so that feature maps are extracted. Instead of using a sliding window over a pyramid of images, or sliding a pyramid of filters to predict regions, RPN introduces *anchor boxes* with different aspect ratios and scales that are slided over extracted convolutional features. The anchor boxes serve as references at multiple scales and aspect ratios. For each location, k anchor boxes are used, with three aspect ratios and three scales, yielding $k = 9$ anchors at every sliding position. A *class layer* outputs $2k$ scores whether there is object or not for k anchors. Another layer called *regression layer* outputs $4k$ values for the anchor box coordinates.

2.5.2 Mask R-CNN

Mask R-CNN is an extension to Faster R-CNN where a segmentation branch is added to predict a binary mask for each object in a region of interest (RoI). The goal is to classify each pixel into a set of categories without differentiating object instances.

It uses the same two-stage architecture of Faster R-CNN [51], while keeping the initial step - region proposal network (RPN) - the same and adds a mask detection branch parallel to object classification and bounding box detection branches. A multi-task loss is defined during training on each sampled RoI as $L = L_{cls} + L_{box} + L_{mask}$. The loss of classification L_{cls} and L_{box} are the same as [51]. The mask loss L_{mask} encodes the evidence of each pixel belonging to each class in the RoI, creating a Km^2 dimensional output. Each RoI has the same dimensions $m \times m$ and K denotes the number classes. The per-pixel level sigmoid function is applied to each mask RoI, and L_{mask} is defined as average binary cross-entropy loss. Among these K number of losses, only one contributes to L_{mask}, which is the k-th mask having the ground-truth label (other mask outputs are ignored).

This definition of loss for masks enables the network to generate each class without competition among classes. The output mask to be activated is determined according to the

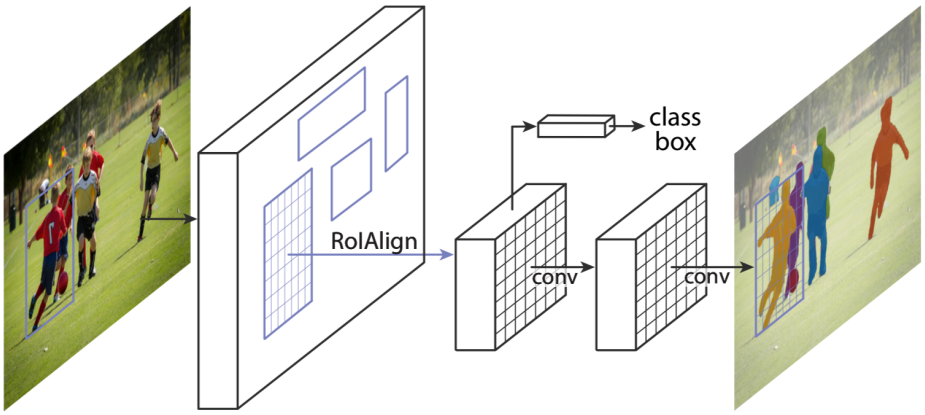


FIGURE 2.8: Mask R-CNN [1] framework for instance segmentation.

classification result of that RoI. Therefore, mask and class prediction operations are decoupled.

3. RELATED WORK

Understanding human activities in the computer vision community have been a hot topic since the late 90s. Aggarwal and Ryoo in [52] classify activities into four categories by complexity as gestures, actions, interactions (human-object or human-human), and group activities. Gestures are meant to be atomic limb movements of an individual, such as *raising a hand*. Actions are performed by a person with the composition of gestures. For example, a running person usually moves both feet and arms back and forth. Interactions can be a human-object or human-human interaction where two or more objects and humans are involved. For example, *person riding a bike* is an human-object interaction and *two persons kissing* is an human-human interaction. Group activities, on the other hand, stand for contextual group activities such as two groups fighting. The definition of Aggarwal and Ryoo covers almost all possible activities. However, the interaction term may be interpreted differently, such as Yang *et al.* [38] define the interactions as touch codes, called proxemics, which aims to find physical relations of people such as hand touches hand or shoulder touches the shoulder. [37, 53] defines the problem as semantic interactions such as a family, group, crowd, etc. photos.

Most of activity recognition works are based on video data since the full behavior is observable in videos. The temporal information has a significant advantage in understanding activities. In images, however, there are less researches than video since there is only a glimpse of the scene, and without the temporal information recognizing an activity becomes a challenging problem. The difficulty is also proportional to the complexity; it is much easier to infer a gesture than a group activity in a still image.

Some researches, on the other hand, tries to take advantage of depth information [54, 55]. Especially with the launch of the Microsoft Kinect sensor, depth data became easy to access [56, 57]. Depth data usually expressed with a 4th channel as *RGB-D* where *D* stands for depth. The most important contribution of depth is that it enables one to isolate and localize the activity from its background.

Another aspect of recognition would be based on the learning method since the computer vision stepped into a new age with the rise of deep learning. Therefore, it would be fair to split to activity recognition methods into two, such as deep and shallow approaches. The

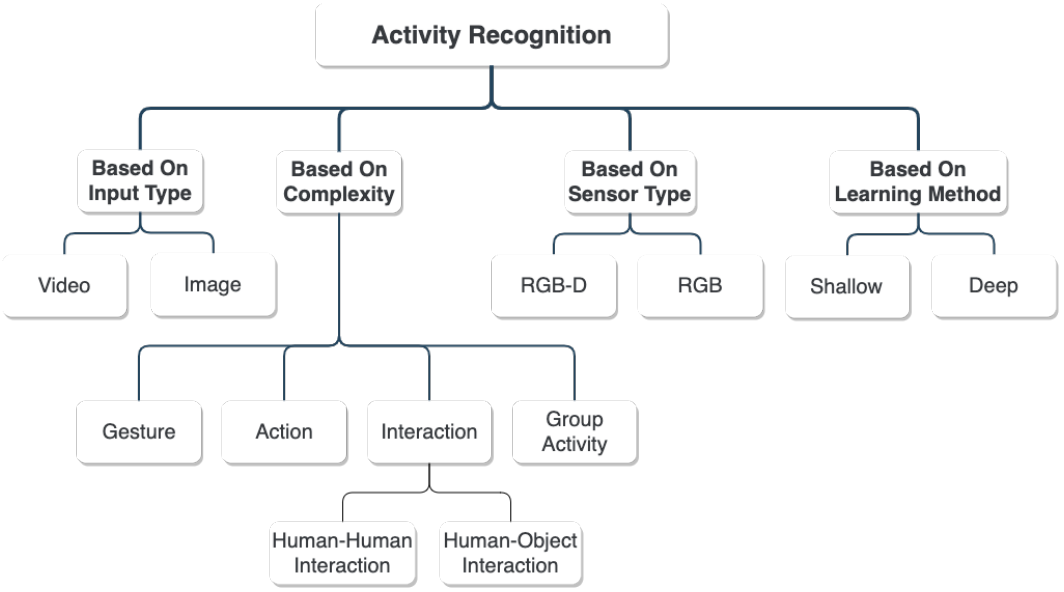


FIGURE 3.1: A taxonomy of activity recognition approaches depending on complexity, input, sensor and learning approach types.

term *shallow* is used in [8] to express the traditional hand-crafted feature engineering-based approaches.

The aspects of recognizing human activities are summarized above. A generalized taxonomy of these aspects is given in Figure 3.1. The hierarchy under *Shallow* and *Deep* learning-based methods could be elaborated a lot deeper. Nevertheless, it is hard to create a general grouping that is also compatible with all video and image inputs or sensor types. For example, a deeper taxonomy for human-human interaction recognition in videos based on learning methods is given in Figure 3.3.

In the second part of this thesis, instance segmentation is used to attend the human body poses to recognize human interactions better visually. (For a recent review on image segmentation, the reader is referred to [58]). Segmentation is one of the hottest subjects of computer vision and has been evolving since the last decade. It moved from detecting visually consistent pixels to semantic segmentation [59] and instance segmentation [1, 60].

Our work is based on recognizing human-human interaction from still images. To have a general idea of our work, a brief review of activity recognition frameworks based on the learning methods are given below. The approaches get complicated from gestures to group

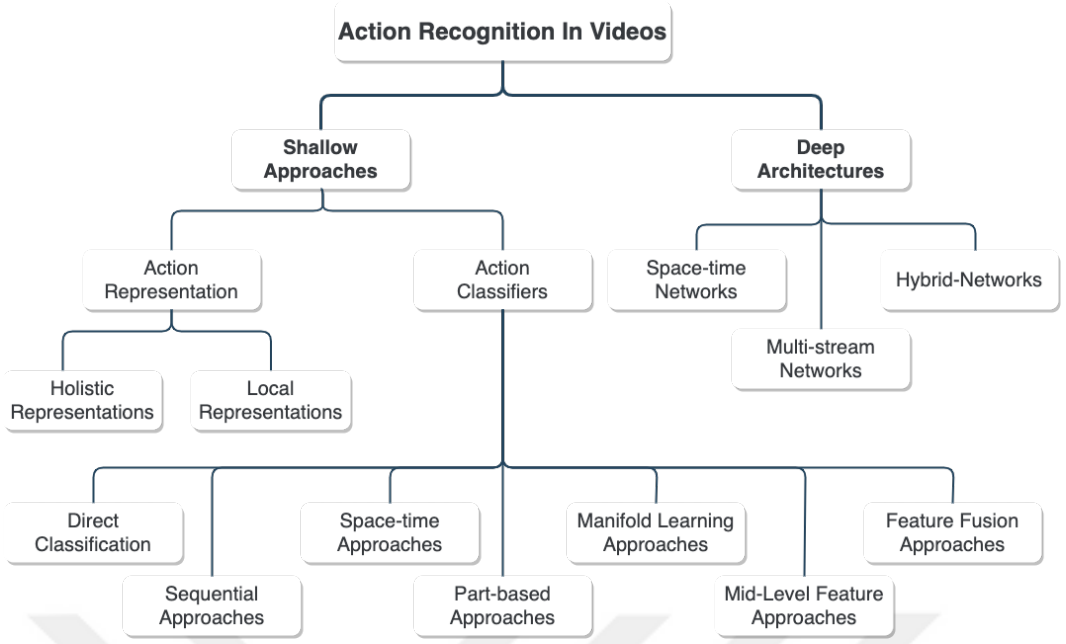


FIGURE 3.2: Summary of action recognition methods in videos given by [8].

activities. However, they resemble each other more or less except that they are vastly separated by the input type -image or video- they use. Therefore we divide the approaches based on input type first and give further details on each separately.

3.1 Recognizing Human Activities In Videos

Recognizing human activities in videos is very popular research in the computer vision community. There are even a lot of surveys on this. Some examine the whole activities [61, 62, 63], most of them examine actions [8, 64, 65, 66] and thereon on human-human interactions [9]. There is also a book on activities [67].

A recent survey on human actions [8] gives a detailed summary of methods on recognizing interactions. Although it is mostly focused on actions, the taxonomy is quite comprehensive and expressive that is applicable for all kinds of activity recognition tasks. Figure 3.2 shows a visualization of the taxonomy given in [8].

Stergiou and Poppe [9] published a comprehensive survey on recognizing human-human interaction in videos. The main focus of the survey is recent techniques that are based on deep learning. A summarized taxonomy based on the learning methods presented in the survey is shown in Figure 3.3. We stick to the original classification given in [9], which

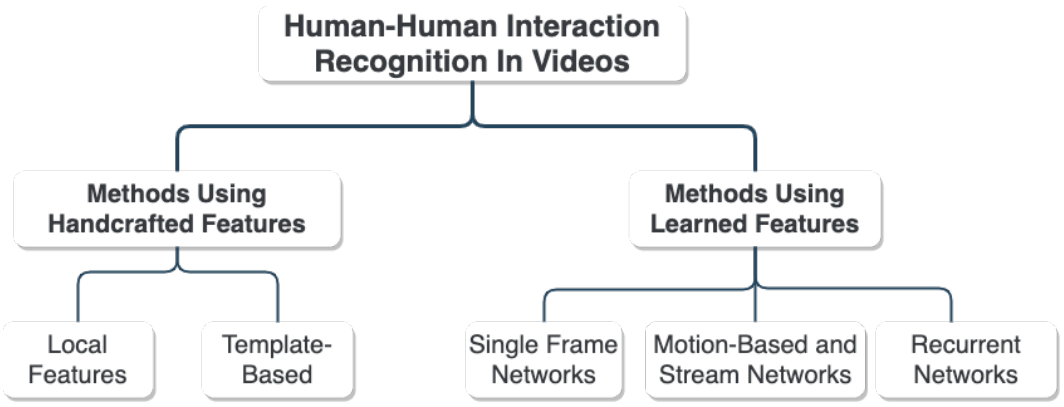


FIGURE 3.3: A taxonomy of recognizing human-human interactions in videos based on the methods given by a recent survey [9].

classifies the studies based on learning methods as *methods using hand-crafted features* and *methods using learned features*. This is actually the same classification as the taxonomy *based on learning method* given in Figure 3.1, where *handcrafted features* corresponds to *shallow* methods and *learned features* corresponds to *deep* methods.

Both of the surveys [8, 9] give a comprehensive intuition and history on activity recognition despite the fact that one focuses on actions [8] and other on human-human interactions [9]. This is mainly because both activities share common interests as recognizing human behaviors spanned through some time. However, we find that the examination of [8] is more detailed on shallow approaches. Thus, we follow the same classification to give a summary of the shallow approaches based on the work in [8].

3.1.1 Shallow Approaches

Shallow approaches or hand-crafted feature engineering-based methods could be divided into two main parts [8, 9, 33, 64] as *representation* and *classification*. In the first part, descriptive features are extracted from frames locally or globally, and these features aggregated from important or all frames of the video. In the second part, these features are used to infer an activity label to tell which class the video belongs to. The following two subsections summarize the representation and classification methods.

3.1.1.1 Activity Representation

For the shallow approaches, it is crucial to represent activities as machine-understandable -in other words, discriminative- feature vectors. The nature of the problem makes it very complicated, such as camera viewpoints, positions, qualities, lighting, environment, occlusion, etc. Features should be expressive enough to solve a reasonable amount of these problems. A rough classification of representation methods would be splitting them into holistic and local approaches [8].

Holistic representations tries to model the entire human subjects in a video. Although they provide a rich representation, they get severely affected by noise and background. For example, Bobick *et al.* [68] presents Motion History Image and Motion Energy Image to create a single image representation of the human action.

Local representations generally find interest points in the video and combines these features with a bottom fashion to create a whole representation. There are several ways of detecting interest points. The most famous ones are based on Harris corners [69, 70], SIFT [71], space-time interest points (STIPs) [72], motion trajectory [73], histogram of gradients (HOG) [41] or histogram of optical flows (HOF) [74]. STIPs provide a good representation of motion, yet it is a short time of period. Therefore some methods track features such as SIFT [71] and HOGs [73].

3.1.1.2 Activity Classification

The action classification step takes the features calculated in the first part and tries to find the most suitable label for the given action. As stated by [8], the classification methods could be categorized into several categories, as given below.

Direct classification methods represent the videos as a feature vector and run a classification algorithm such as SVMs [75]. This method is not entirely successful as it does not incorporate the temporal information. *Sequential approaches* tries to solve this issue by modeling the temporal movement of the features as hidden Markov models (HMM) [76] or conditional random fields (CRFs) [77]. *Space-time approaches*, on the other hand, consider the correlation between feature points through space and time [78, 79].

Part-based approaches detects human body parts in frames and models the activity with structural information collected by these body parts [80, 81]. A different line of work is *manifold learning approaches* used in action recognition [82]. The aim is to reduce dimensions of human silhouettes over time to small feature manifolds. *Mid-level feature approaches* propose to express features with a higher level, such as Bag-of-words [83, 84]. Finally, *feature fusion approaches* tries to combine different features to create a better representation of the activity [85, 86].

3.1.2 Deep Architectures

Shallow approaches are usually created based on hand-crafted features. These features require domain-specific engineering that prevents a generalized method. Deep architectures eliminate the sub-optimal feature selection process by selecting the best parameters among millions. However, straightforward deep architectures generally take an input image and generate a classification score. This is a problem when recognizing human activities because temporal information is a critical cue. Kong and Fu [8] categorize the solutions to exploiting temporal information into three: Space-time, Multi-stream and Hybrid networks.

Space-time Networks uses 3D convolutions to capture the temporal information. 3D convolutions are first presented in [10]. Multiple sequential frames are stacked together to form a cube, and a 3D convolution kernel is applied to this cube. Tran *et al.* [87] presents a network architecture based on 3D convolutions called C3D. It is shown that 3D convolutions learn better representations of spatio-temporal features than 2D convolutions.

Multi-stream Networks proposes to use multiple convolutional networks to exploit both temporal and spatial information. Karpathy *et al.* [13] proposes several variations of 2D convolutions applied to consecutive frames that are fused at different levels (layers) of the network. Three network variations based on the fusion technique are presented: *early*, *late*, and slow fusion. Early and slow fusion use 3d convolutions as in Space-time Networks. Late fusion, on the other hand, fuses the output of two single networks in the first fully connected layer at the end. Input to each network is given as two frames in a video with a distance of 15 frames. Simonyan and Zisserman [6] proposes to use a two-stream architecture where the optical flow is used as the temporal information. Softmax outputs of each stream are fused to generate a single classification score. [88] extends the work of [6] by using it as a feature extractor centered around the detected point trajectories. A fisher vector representation of these features is created and fed to linear SVM to train the action classifier. Girdhar *et al.*

[89] proposes a new pooling layer called ActionVLAD that pools features from both spatial and temporal domains. ActionVLAD is used in a two-stream fashion to exploit motion information better.

Hybrid Networks use recurrent neural networks(RNN) to exploit temporal information. Such networks use the standard CNN architectures as a feature extractor for each frame and feed the features into RNNs, hence the name Hybrid Networks. Plain RNNs suffer from vanishing gradient problem. Therefore improvements such as *Long short-term memory (LSTM)*[90] is used to encode temporal information. A typical example of Hybrid Networks is presented in [11]. Features from single frames extracted by CNNs are fed to LSTMs to learn the temporal nature of the action. The average of each LSTM node is used as the final classification score of the video.

Deep neural network architectures to classify human activities are summarized in this section. Figure 3.4 demonstrates a visual summary of these architectures.

3.2 Recognizing Human Activities in Still Images

3.2.1 Action Recognition

Human action recognition is widely studied in still images. One of the earliest methods [91], hand-labels joints of human limbs to create 3D representations of poses and uses the distance of endpoints as features to classify actions. Wang *et al.* [92] use similarity of edges to cluster human actions in sports news. [93] combines the existence of certain poses, scenes, and objects. [94] represents human poses with histograms of rectangular and orientational regions, then classify with Support Vector Machines (SVM). [95] presents Pose Primitives, which are based on Histogram of Oriented Gradient (HOG) [41] features. [96] combines features of person and background. Yao *et al.* [97] defines attributes for verbs, which describe human action, such as sitting and riding, parts for poselets, and objects related to the actions. They also introduce the well known *Stanford 40 Action Dataset* with this work. [98] combines pose and context features together and train an SVM classifier. [99] defines a graph model encoding spatial configuration of body parts. [100] investigates relations between actions and objects, which are usually around when a specific action is being performed. A recent work [101] introduces quite a challenging dataset called MPII Human Pose Dataset (MPII), which contains 393 action classes and 15205 images in total.

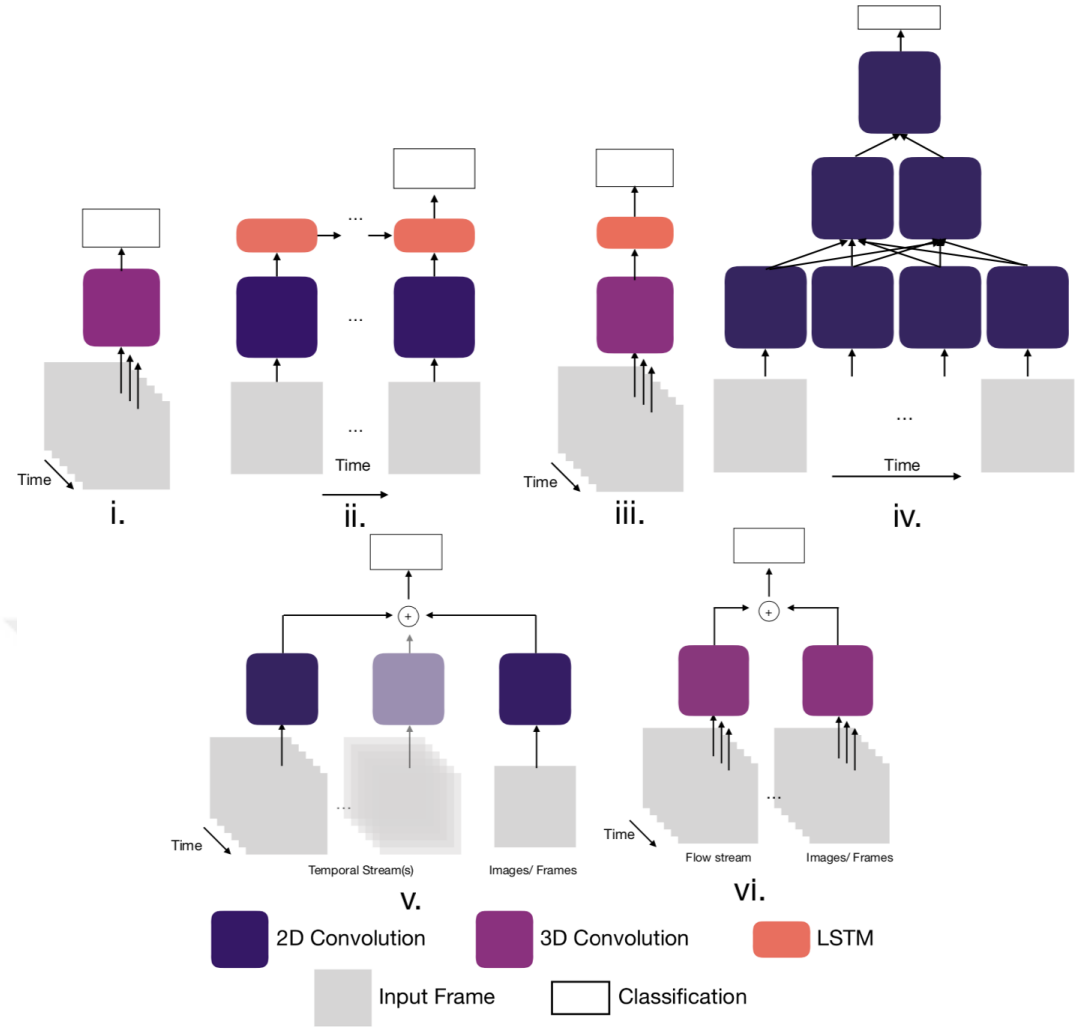


FIGURE 3.4: Video classification networks. (i) 3D-convolution [10], (ii) 2D-Convolutional LSTM over a sequence of frames [11], (iii) 3D LSTM [12], (iv) slow-fusion [13], (v) two/multi-stream CNN [6, 14] and (vi) two-stream 3D-Conv network [15]. This figure is taken from [9].

[102] exploits the contextual cues around the person taking the action. A modification is applied to R-CNN [103] to use more than one region while still being able to localize the action. [104] proposes an attentional pooling method that replaces the last convolutional and fully connected layers. Attention-based approaches have been a topic of interest that is explored in many recent works. Minaee and Abdolrashidi [105] propose to use an attentional convolutional neural network for recognizing facial expressions. Wang *et al.* [106] proposes a framework that combines residual networks with multiple attention modules. [107] proposes an Attention Branch Network which uses attention maps for a visual explanation of images.

Lu *et al* [108] introduces a different approach to learn human actions. By adding a person detection branch to the model, it is also forced to find the pixels that belong to a human body. The loss of two branches is combined to create a single end-to-end learning task. Therefore, the layers before the branch have to learn features that represent both actions and human pose at the same time. This method could be considered as the opposite of the proposed approach, which uses the already detected poses as input to the learning framework.

3.2.2 Human-Object Interaction Recognition

Recognizing human-object interactions on still images is another topic of interest. [109] uses the interaction between humans and objects to detect them in the images better. Yao and Li [25] describe a new image feature representation called *grouplet*, which is a structural representation of discriminative visual features and their spatial configurations. A new dataset called People-Playing-Musical-Instruments (PPMI) is introduced for evaluation. [26] combines low-level pixel features and high-level structural features based on body parts. A recent work introduces a challenging dataset HICO [110] for human-object interactions. HICO contains 600 interactions which cover 117 actions with 80 common objects. For each image, all salient interactions are labeled as ground-truth; in other words, it is a multi-labeled dataset.

3.2.3 Human-Human Interaction Recognition

Human-human interactions in images is still a less-studied subject that is open for further progress. Yang *et al.* [38] define the interactions as touch codes, called proxemics, which aims to find physical relations of people such as hand touches hand or shoulder touches the shoulder. [37, 53] defines the problem as semantic interactions such as a family, group, crowd, etc. photos. On the other hand, [37] and [53] combines several features such as distance between people, people configuration, camera pose, shot position, and the number of group layers. A recent work [111] uses a deep learning framework and a multi-stream fusion scheme to recognize human events. Events are broader than interactions since they involve any activity that one or more people perform.

Another study area that could be related to our work is event recognition in still images [112, 113, 114]. Event recognition research aims to recognize a certain scene or event in images or videos. Datasets in this field are different from ours. Event recognition datasets describe an event like Christmas, weddings, etc. In such images, the main focus is not the people, but visual elements for an event. In a multi-person interaction recognition problem, we focus on the presence of people and try to infer the interaction based on images of people.

Convolutional Neural Networks outperforms traditional classification methods. Although it's straightforward to process a single input or feature such as an image, the problem of integrating information from different inputs remains to be solved. Early studies consider combining inputs directly at the pixel level [13], or training separate networks and fuse the classification scores [6]. On the other hand, later studies [7, 50] show that fusing at a point before the classification layer enables the networks to be trained end-to-end, which produces better success rates. Inspired by these works, we define several pose representations to be processed as independent inputs to CNNs. Furthermore, to combine the information obtained from these inputs, we fuse the CNNs similar to the approach of [7, 50].

The first part of our work uses facial features to recognize human interactions. In this work, we are inspired by [27], which uses face detection responses to recognize social human interactions in video sequences from a first-person perspective camera. They propose to use Markov Random Field for frame-based feature representations and a Hidden Conditional Random Field to represent sequence-based features. In our work, we propose several simple features based on face regions for recognizing human-human interactions in the images.

In the second part, we study the effects of poses on recognizing human-human interactions in images. Our approach differs from the related work as we aim to find high-level social interactions such as hugging or handshaking. Our work validates that the human pose information possesses a large degree of discriminative power in recognizing human-human interactions.

4. FACIAL DESCRIPTORS FOR HUMAN INTERACTION RECOGNITION IN STILL IMAGES

In this chapter, we describe the facial descriptors and the learning procedure that we have proposed for interaction recognition.

Our approach begins with the detection of the faces. For this purpose, we first apply a recent algorithm of [2], since it outputs three essential information about the faces: (1) Orientation of the face in the range of [-90°, 90°] with a resolution of 15°, (2) location of the face in the image and (3) size of the face in pixels. The orientation of a face is defined as the angle of the face with respect to the imaginary axis that crosses from the midpoint of the chin and the forehead. For reducing the number of false negatives in face detection, we also employ the OpenCV implementation of [115], which only gives the location and size of the images and whether they are frontal or profile. The face detections from these two approaches are combined in the following way: (1) If both of the detectors find a face in the same region, [2]'s output is used, since it is shown to be more accurate and it outputs face orientation estimates as well as face locations. (2) While using the Viola-Jones face detector, if the only frontal face is detected, the orientation is assumed to be 0°. If a profile face is detected, then the orientation is assumed to be 90°. If both frontal and profile face detectors fire within the same region, it means that the orientation of the face is between [0°, +/-90°]. To quantize the angle, the intersection ratio is normalized to [0°-90°] interval.

After detecting the faces, we extract several mid-level descriptors based on the facial regions. Below, we introduce each of these descriptors.

Histogram of Face Orientations (HFO): In order to account for the distribution of face orientations, we propose to use Histogram of Face Orientations (HFO), which simply is based on the count of face orientations for each angle in an image. In other words, it is the distribution of face orientation frequencies in an image. This descriptor has 13 feature dimensions (13 histogram bins), which corresponds to 15°resolution in [-90°- 90°] interval. Figure 4.1 shows some example HFO descriptors.

Histogram of Face Directions (HFD): We observe that using orientations with a lower resolution can also be useful to discriminate interactions. Based on this observation, we

FIGURE 4.1: Extraction of the histogram for the orientation of faces. First row shows face detection and orientation estimation results and second row shows the histograms of face orientations (HFO).

form a coarser histogram representation of the face orientations and call it Histogram of Face Directions (HFD). This HFD feature comprises the distribution of direction frequencies in the images. Directions are defined as left, front and right which correspond to the angle intervals [-90°, -45°], [-30°, 30°] and [45°, 90°] respectively. This descriptor is basically a coarser form of HFO, and it has three dimensions.

Distances of Faces (DF): The relative locations of the faces can also be representative for an interaction. In order to capture this information, we propose to use histogram of relative distances between faces in terms of pixels. Let $L_i = (x_i, y_i)$ be the location of each face center and D_i is the distance of each face to the global center of faces. If there are two faces in the image, the distance is simply the Euclidean distance of the faces. If there are more than two faces in the image, first the center point C of the faces is calculated

$$C(x, y) = (\frac{1}{N}\sum_{i=1}^{N} L_x, \frac{1}{N}\sum_{i=1}^{N} L_y) \tag{4}$$

where N is the number of faces detected in the image. To normalize the distances, we divide each of the distances to the maximum face size $S = max(s_i)$, where $s_i = max(s_x, s_y)$ is the max edge size of ith face. Then, D_i is calculated as

$$D_i = \min(K, \max(1, \lceil \sqrt{|L_i - C|^2}/S \rceil)) \tag{5}$$

where, K is the maximum distance value for a face.

We then form the histogram of distances, which simply comprises the distribution of distance frequencies in the images. The maximum distance value K defines the number of bins of the histogram since distance values are bounded to the interval $[1, K]$ and discretized into

equal-sized bins. We set K such that at least 95% of the faces are covered within the region defined by this maximum distance value, and in our dataset, we observe that $K = 5$ satisfies this condition.

Circular Histogram of Face Locations (CHFL): In order to capture the relative layout of people within an image, we propose to use a histogram of their locations. For this purpose, a circle is fit to the center of the extracted faces within an image. The center C of the faces is calculated as in Eq. 4. Radius r of the circle is calculated as the maximum face distance to the center: $r = max(D_i)$. Thereafter, the circle is divided into equal parts like a pie chart by an angle α. Then the distribution of faces over the pies is calculated as a histogram. To decide which face lies on which pie, the angle ϕ of the line that crosses both C and L_i is calculated as:

$$\phi = \arctan\left(\frac{y_i - C_y}{x_i - C_x}\right) \tag{6}$$

where x_i and y_i are the x-y coordinates of ith face in the image. having identified the pie index of the face in the circle, the last step is to create a histogram that comprises the distribution of face frequencies over the pies of the circle. The histogram consists of $2\pi/\alpha$ bins, and the location of the face projected on the histogram as α/ϕ. Figure 4.2 shows example cases on how this CHFL global descriptor is extracted.

Grid Histogram of Face Locations (GHFL): Similar to CHFL, we also form the grid histogram of face locations, in order to capture the spatial layout of the multiple people within an image. For this purpose, the image is split into the $M \times N$ size grids where the center of the middle grid is the center of the faces. A $M \times N$-bin histogram is created, and the number of faces falling within the respective grid is computed. Our preliminary results show that 1×3 grid size gives the best results since, for most of the interaction images, faces lie on the horizontal plane.

In addition to the visual features described above, we also include the number of different facial orientation directions to our list of global facial descriptors. Our final feature length for the combined descriptor, therefore, becomes 31.

Scene features: In order to capture the general characteristics of the scene and to investigate its influence on human-human interaction recognition, we use GIST descriptors proposed by

[42]. GIST features provide a low-dimensional representation of the global scene layout.

In addition, we utilize Bag-of-Words (BoW), both as a baseline and a global scene descriptor. We extract dense SIFT features from the images and create a codebook using k-means (k=1000). Then, each image is represented with the histogram of codewords. We additionally employ spatial pyramid matching (SPM) of [44], with 2×2 grids.

Deep features: State-of-the-art in many computer vision classification tasks has received a significant gain in performance with the introduction of deep learning and the Convolutional Neural Network (CNN) architectures.

In order to explore the effect of these architectures over the human interaction image recognition, we use two kinds of deep features, which are trained using the large scene dataset called Places ([116]). This deep feature is called Places-CNN and is trained using the Caffe package over the dataset of 2,448,873 images of 205 categories. The output of the FC7 layer, which is 4096-dimensional, is taken as the deep feature, and a linear SVM is trained using our training set.

We also evaluate a similar feature called Hybrid-CNN, based on the same architecture, but trained over the larger dataset that combines both the Places and the ImageNet datasets ([116]). This joint training set has 3.5 million images from 1183 categories. Due to the broadness of classes in the training set, this Hybrid-CNN feature is likely to capture both the scene and the object characteristics. The reason for not using ImageNet alone is that ImageNet is mostly oriented towards singular object information. By using Places dataset together with the ImageNet dataset, we aim at capturing scene information, as well as to object information. The results of using both of these deep features are given in the Experimental Evaluation section.

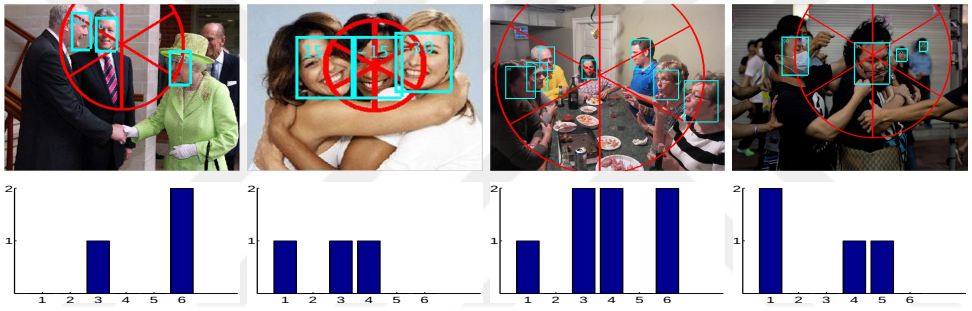


FIGURE 4.2: Example images for the calculation of Circular Histogram of Face Locations(CHFL). A circle is fit to the center of the faces within the image, with a radius equal to the maximum distance of a face from the center. The circle is split into 60°pie regions and the distribution of faces falling into the (360°/60°) pies are calculated via a histogram.

5. MULTI-STREAM POSE CONVOLUTIONAL NEURAL NETWORKS FOR HUMAN INTERACTION RECOGNITION IN IMAGES

The human pose is one of the most distinctive features for recognizing human-human interactions. In order to make use of most of the pose information, several different representations of poses, and their utilization within CNN architectures are explored. In accordance with this purpose, three representation forms along with the original RGB inputs are employed: (i) pose masks (PM) images to describe poses for interactions (ii) pose only (PO) images to estimate context information (iii) and finally pose highlighted (PH) images are introduced to elevate background context while preserving pose information for interactions. The other main component of the proposed approach is to predict an interaction label for the given inputs. To this end, multi-stream pose CNN architectures are introduced that aggregate the information from multiple streams and predict an interaction label directly from the given inputs for each. Below, each of these representations and the corresponding architectures is described in detail.

5.1 Pose Masks

Since by intuition, poses can be considered as the key elements of an interaction, we aim to evaluate this property and assess how well the poses describe an interaction. In accordance with this purpose, we propose to use *Pose Masks Images* (PM). To this end, we first detect the poses within the images using MaskRCNN [1], where corresponding response maps yield individual poses. More formally, let P_i be a binary image with a size equal to the original image for each pose detected by [1]. The pixel $P_i(x, y)$ is 1, if it is a pose region and $P_i(x, y) = 0$ otherwise. Next, the combined pose mask image PM is defined as the combination of these individual poses and is calculated as follows:

$$M_i(x, y) = max(M_{i-1}(x, y), P_i(x, y)) \tag{7}$$

where M_i is the combined binary image of i pose maps, and $M_0(x, y) = P_0(x, y)$. After this combined pose map is formed by combining all the pose detections, it is converted to an RGB image with background areas set to a constant value μ:

$$PM(x, y) = \begin{cases} M(x, y) \times 255, & \text{if } M(x, y) = 1 \\ \mu, & \text{otherwise} \end{cases} \tag{8}$$

We select μ as mean RGB values of ImageNet [45]. The example Pose Mask images are shown in Figure 1.2 column b. The effect of μ value is also observable in the figure, see how the background areas of the images are greyed out.

5.2 Pose Only Images

Pose Mask Images are capable of describing the poses, yet they suffer from the lack of context. To cope with this, we also propose to use an alternative representation, which we refer to as Pose Only (PO) Images. In this representation, instead of using the mask of the poses, the regions containing the poses are used as is, while the background is set to a constant value μ. More formally, the Pose Only (PO) Image of an image $I(x, y)$ is calculated as:

$$PO(x,y) = \begin{cases} I(x,y), & \text{if } M(x,y) = 1 \\ \mu, & \text{otherwise} \end{cases} \tag{9}$$

M is the combined binary mask image defined in Equation 7. Example PO images are shown in Figure 1.2. As can be seen, these images include the regions of the images that are masked by the pose detections.

5.3 Pose Highlighted Images

PO Images tell a lot about the interaction. For instance the people in the middle row of Figure 1.2 column **c** are clearly *talking* to each other. However, in some cases PO fails because of two main reasons: background context might be missed and pose detection does not give perfect results. Although MaskRCNN [1] does a great job detecting people in images, it's not 100% accurate in pose estimation. There are a lot of cases where some of the limbs are either incomplete or not even detected. An example case is shown in Figure 1.2 last row, where the face of the man is missed by the detection algorithm. To decrease the effect of these problems, we introduce Pose Highlighted (PH) images. In PH images, a multi-Gaussian shaped mask, created from each pose mask, is multiplied with the original image to increase the contribution of human containing areas; while decreasing the effect of background areas. PH contains only a small amount of background context information. However, it recovers the artifacts of pose detection and creates more natural pose images to process.

Let $F = [F_x, F_y]$ represents the maximum size (length and width) of a detected pose in an image, binary pose mask M is smoothed by a Gaussian filter G as:

$$S = M * G \tag{10}$$

where M is the pose mask given in Equation 7 and $*$ denotes convolution operation. Size of the filter G is directly proportional to F as $G_{size} = \theta \times F$, where θ is a constant. To make sure that filter sizes are always reasonable, we limit the size with lower (G_{min}) and upper (G_{max}) bounds where $G_{min} \leq G_{size} \leq G_{max}$. In our experiments we empirically determine θ as 0.2, G_{min} as $[25, 25]$, and G_{max} as $[50, 50]$. Although this makes the pose regions diffuse outside, smoothing causes loss of too much weight near edges of pose regions. To avoid this, smoothed S and original M masks are combined with an *alpha* filter as follows:

$$S'(x, y) = M(x, y) \times \alpha + S(x, y) \times (1 - \alpha) \tag{11}$$

where $\alpha = 0.05$ is selected empirically. This restores the information around pose edges, yet it needs to be normalized to $[0 - 1]$ interval to be applied to the image:

$$S''(x, y) = \frac{S'(x, y) - min(S')}{max(S') - min(S')} \tag{12}$$

Finally, the pose highlighted image *PH* is created as:

$$PH(x, y) = I(x, y) \times S''(x, y) \tag{13}$$

Figure 5.1 visualizes this formation process of PH Images. Three samples are shown in the figure. It can be seen that some of the limbs are recovered to a certain extent.

5.4 Multi-Stream Pose CNNs

5.4.1 Multi-stream CNNs

Poses may not cover all the visual information for recognition of interactions, since additional contextual content, such as the scene and the overall layout may also involve significant cues for recognition. In this regard, poses can be used as an auxiliary contributor to the visual information extracted from the overall scene. Stemming from this point of view, we introduce the "multi-stream pose Convolutional Neural Networks".

FIGURE 5.1: Process of creating a pose highlighted image. Rows are from top to bottom; **1.** Original images. **2.** Pose mask image detected from original image. **3.** Smoothed pose mask. **4.** Pose and smoothed masks combined with *alpha* filter where $\alpha = 0.05$. **5.** Original Image. **6.** Pose Only (PO) image, included here to compare with pose highlighted image. **6.** Pose Highlighted image generated by multipliying **c** with original image.

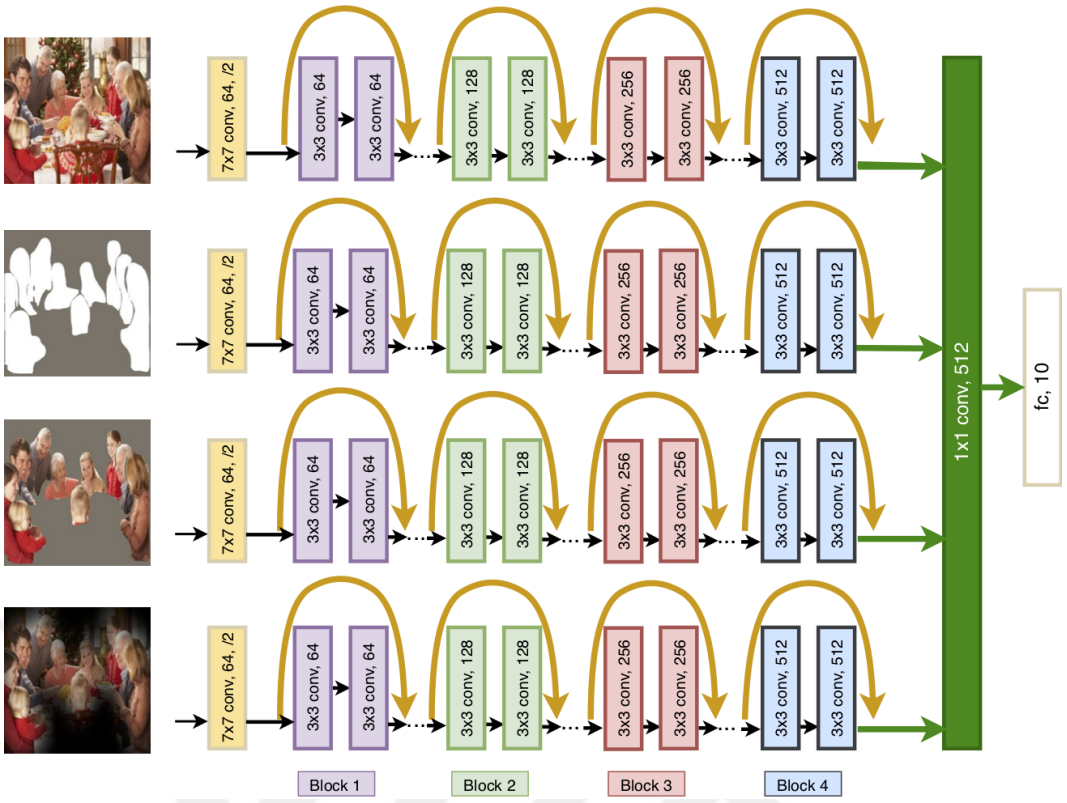


FIGURE 5.2: Demonstration of our 4-Stream Pose Network based on ResNet152 model. Each row represents a stream in the overall model. From top to bottom, streams are RGB, Pose Mask, Pose Only and Pose Highlighted. Fusion is done right before the final FC layer with a 1x1 Convolution Block.

The overall idea is based on the seminal work of Simonyan and Zisserman [6], where two different Convolutional Neural Networks that operate on different input types (RGB and Optical Flow) are used in conjunction for action recognition in videos. Simonyan and Zisserman were the first ones to introduce the idea of using two CNNs operating on different input forms. In our context, each stream corresponds to a CNN that operates on different inputs, which are RGB, Pose Masks (PM), Pose Only (PO) and Pose Highlighted (PH) Images.

Specifically, given the raw RGB image, PM, PO and PH images are extracted, and each of these meta-representations is fed into separate CNN streams. We investigate several ways of combining different streams: (1) Two-stream networks where the first input is the raw image, and the second input type is PM, PO or PH images. (2) Three-stream networks that operate on RGB and two of the pose representations. (3) Multi-stream networks that operate on all of the pose representations. Figure 5.2 shows this third model, where each of the individual streams is based on ResNet [4] architecture.

5.4.2 Fusing Streams

In cases where more than one network streams are used for a single learning task, the information from each stream has to be combined together at some point. When the output dimensions of the layer to be fused are equal, it's straightforward to fuse the streams. In these kinds of situations, different fusion techniques are considered such as sum, max, concatenation, convolutional layer, etc. In our preliminary experiments, we have observed that fusion with convolution produces the most stable results. This finding is also in accordance with the findings of Feichtenhofer *et al.* [7] which shows that fusing by a convolutional layer is a good choice. In order to do this fusion, basically, features from each stream are vectorized and concatenated as:

$$y^{cat}_{N,Kd,1,1} = f^{cat}(x^1_{N,d,1,1}, x^2_{N,d,1,1}, ..., x^K_{N,d,1,1}) \tag{14}$$

where $X^k_{N,d,1,1}$ denotes the feature vector of a stream with size d, batch size N, and k represents the kth stream, where $k \in \{1 \dots K\}$. Next, the concatenated features are fed to a 2D convolution layer with input size $(N, Kd, 1, 1)$ and output size $(N, d, 1, 1)$. While reducing the feature size, the convolution also learns the correspondences of each feature map which makes it a powerful method. In our model, 1x1 convolutional layer is used to fuse multiple streams.

We use cross-entropy loss along with a softmax classifier layer for training the network. The loss function is defined as:

$$L = -\sum_{i=1}^{C} y_i log\hat{y}_i \tag{15}$$

where C is the number of classes, y is a one-hot label vector, and $\hat{y}$ is the vector of class probabilities. Note that in a multi-stream network $\hat{y}$ is function of each input stream formulated as $\hat{y} = f(x^1, x^2, ..., x^K)$.

In our multi-stream architecture, ResNet [4] is used as the primary base CNN architecture within the individual streams. ResNet [4] is one of the most successful recent CNN architectures that offer high accuracy recognition rates for the image classification problem. The idea behind its success is the residual learning architecture, which is made possible by the usage of *skip-connections*. In general, skip-connections refer to those connections skipping one or more layers. In ResNet, the authors simply perform *identity* mapping. More formally, an input x goes through *conv-relu-conv* series. This produces some $F(x)$ which is added to

the original input x, resulting an $H(x) = F(x) + x$. $H(x)$ would be equal to $F(x)$ in a regular CNN without skip-connections. Instead of computing $F(x)$ as a straight mapping from x to $F(x)$, the term to be added to the original x is being computed via the layers. In other words, the residual block is computing a small change to the original input x, making the CNNs more easily trainable by bypassing the vanishing gradient problem. ResNet [4] has four main blocks, which are created by stacking residual blocks, and a classification layer at the end. Simplified versions of these blocks are shown at the bottom of Figure 5.2.

5.4.3 Fusion Architectures

A multi-stream CNN architecture requires the information extracted from each stream to be combined at a point. Early studies such as [6] use late fusion technique where each stream is trained individually and fused by averaging the scores or training a linear classifying on the scores. On the contrary, recent studies such as [7, 50] consider early fusion where the whole architecture is trained end-to-end. We evaluate both early and late fusion in our experiments. On the other hand, early fusion is chosen for benchmarks because it's slightly better in general. Although early fusion is slower at training time, there is no much difference in test time.

As for the late fusion, every single stream's score is summed to obtain the final classification score. For the early fusion case, we opt to fuse every stream at the same location. The reason for this choice is that we use the same network architecture for each stream with different inputs. Hence the model does not get complicated and has a uniform setup. As a consequence, our task is reduced to selecting a common fusion point for all streams.

In the ResNet [4] architecture, which is the primary network architecture in this work, there are 4 main blocks and a classification layer at the end. In our experiments, fusing before each of these blocks and classification layer are tested to find out the best fusion point. Figure 5.2 is an example of fusing before the classification layer. The output of individual streams is concatenated at the fusion layer and fed to a 1x1 Convolutional block. The rest of the network is processed the same as in the original CNN model. Our experiments on ResNet architecture shows that the success of the network increases while the fusion point gets closer to the end of the network.

6. DATASET

Human interaction recognition is a subjective term, especially in the image domain. Images contain only a glimpse of interaction; therefore, it is hard to define solid interaction categories. For example, Yang *et al.* [38] defines interactions as touch codes, such as hand touches shoulder. In this work, interactions are defined as observable behaviors between two or more people. As for the first part, the role of human faces is investigated. Since this is the first research on this basis, a new dataset called Human Interaction Images (HII) is collected. Although HII is a challenging and large enough dataset for a traditional classification method, it is relatively small to train deep learning tasks. Thus, for the second part of our work, HII is extended in the number of instances per class, which is called HIIv2. These two datasets are further elaborated in the following sections.

6.1 Human Interaction Images (HII)

In order to evaluate the proposed facial descriptors and their effects on human-human interaction recognition, we collected a new image dataset that includes ten human interaction classes. These classes are *boxing-punching, dining, handshaking, high five, hugging, kicking, kissing, partying, speech*, and *talking*. Each class contains at least 150 images, forming a total of 1971 images. When collecting the dataset, we gather images such that one of the target interaction classes is present, and at least one person has a visible facial region in each image. The images for the *boxing-punching, handshaking, high five, hugging, kicking, kissing* and *talking* classes usually include two to three people, whereas the number of people in the images for the *dining, party* and *speech* classes varies significantly. Figure 6.1 illustrates example images from this dataset. Most common queries are used to download the images. For example, queries such as *man kissing woman, lovers kissing, mother kissing son, etc.* are used to download images. Different search engines are used to diversify the query results.

HII dataset is used primarily for the evaluation of our work on faces. Our approach uses the size, location, and orientations of faces in the images. Face detectors are used to collect this information automatically. Although the face detectors are capable of detecting the majority of faces, they are not 100% accurate. There are a considerable amount of faces that are not detected or false-positive detections. Therefore, to validate our work on faces, ground-truth face information in the whole dataset is annotated by human annotators. An application is created to annotate images. For each image in the dataset, a rectangle, which corresponds to the size and position of the face, and an angle is determined for each face. Figure 6.2 shows a screenshot from this application. A sample from the *High-Five* class is shown in the figure.

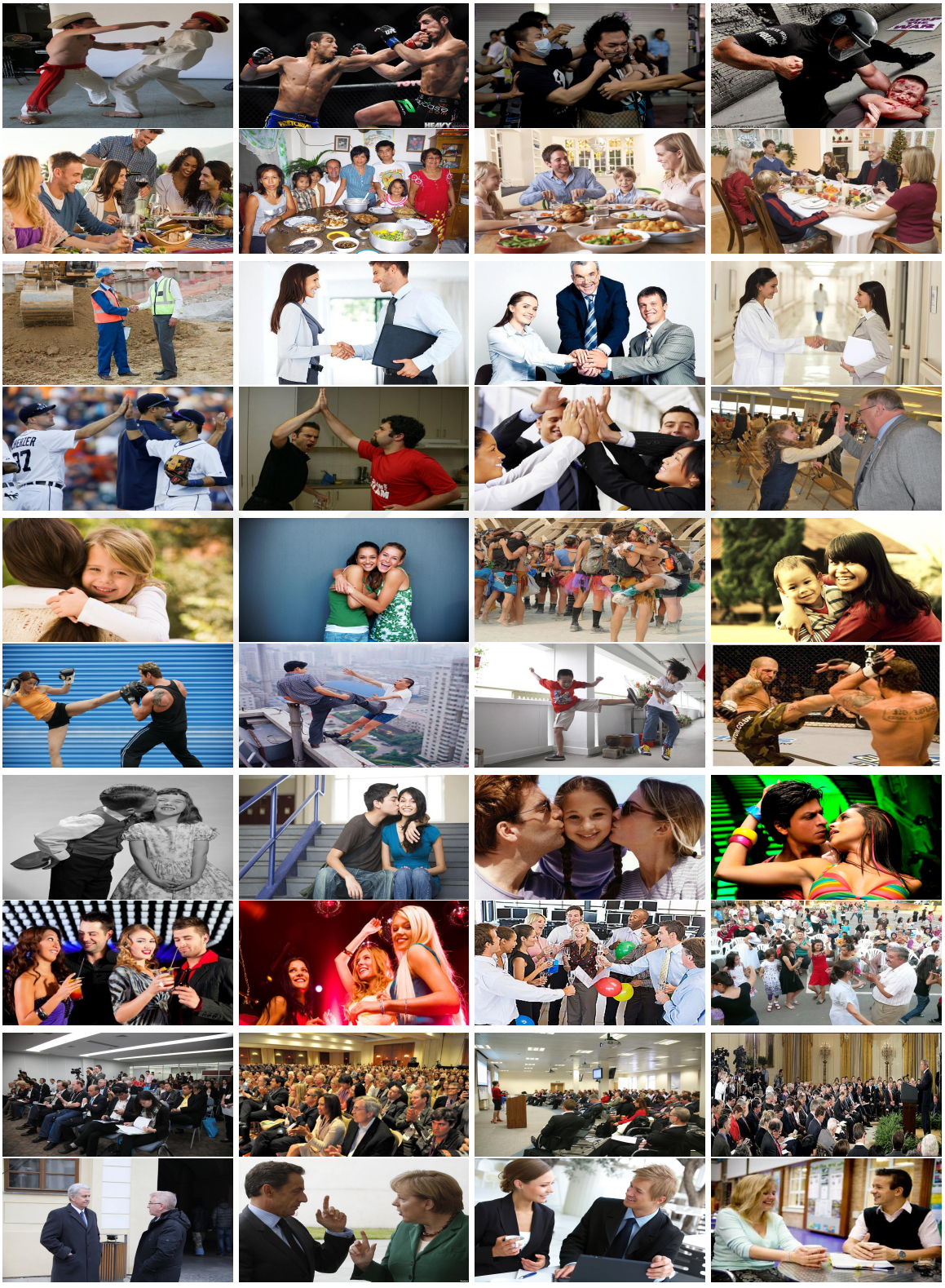


FIGURE 6.1: Example images from HII dataset. Each row belongs to one of the ten classes for this dataset, *boxing-punching, dining, handshaking, highfive, hugging, kicking, kissing, partying, speech* and *talking*. Note that, the poses and appearances of the people in interaction is quite diverse, making it a challenging dataset.

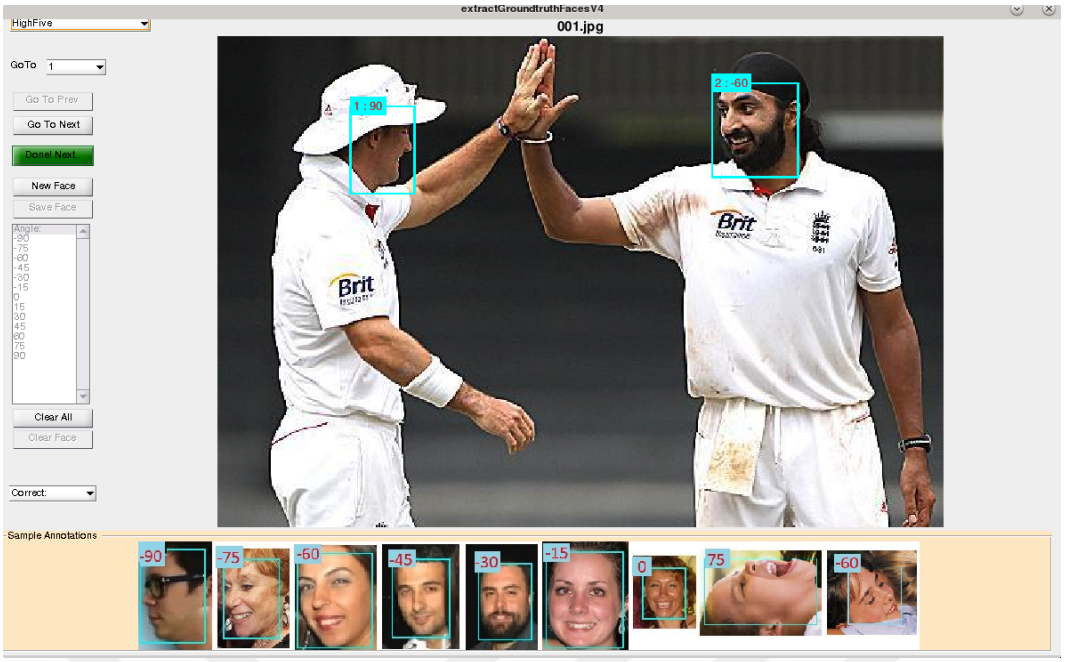


FIGURE 6.2: Example screenshot from the face annotation application. Two faces are annotated with size, location, and orientations. Bottom row shows sample annotations as reference.

Two people are involved in the image, whose faces are clearly visible. Rectangles, covering the faces, are drawn, and the orientations of faces are selected as $90°$ and $-60°$.

HII dataset is available online to the public at `https://gtanisik.github.io/projects/hii` address. The ground-truth annotations are also shared on the web site.

6.2 Human Interaction Images v2 (HIIv2)

With the rise of deep learning, the hunger for more data has increased substantially. Studies [117] have shown that small datasets with images at the scale of the hundreds are insufficient for deeper convolutional neural network architectures to converge. The size and the number of available datasets are also quite limited for human-human interaction recognition in images. To the best of our knowledge, there is only one image dataset available online for human-human interaction recognition that is out HII dataset. This is a relatively small dataset to acquire a well-trained deep CNN architecture. We have first tried to evaluate our models on this dataset; however, during training, we observe perfect overfitting with a deep CNN architecture such as ResNet [4].

To address this issue, we extend the *HII* dataset, such that the number of images in each category is at least 1000. To extend the dataset, queries are created for each class, and they are

FIGURE 6.3: Example images from HIIv2 dataset. Three example images are shown for each of the ten classes for this dataset, *boxing-punching, dining, handshaking, high five, hugging, kicking, kissing, partying, speech* and *talking*. Note that, the poses and appearances of the people in interaction is quite diverse, making it a challenging dataset.

TABLE 6.1: Comparison of the image datasets on human-human interaction recognition.

Dataset	b&p.	din.	h.sh.	h.fv.	hug.	kick.	kiss.	pty.	sp.	tlk.	total
HII	193	207	226	198	213	151	196	193	208	187	1972
HIIv2	1000	1084	1204	1027	1000	1002	1044	1053	1081	1007	10502

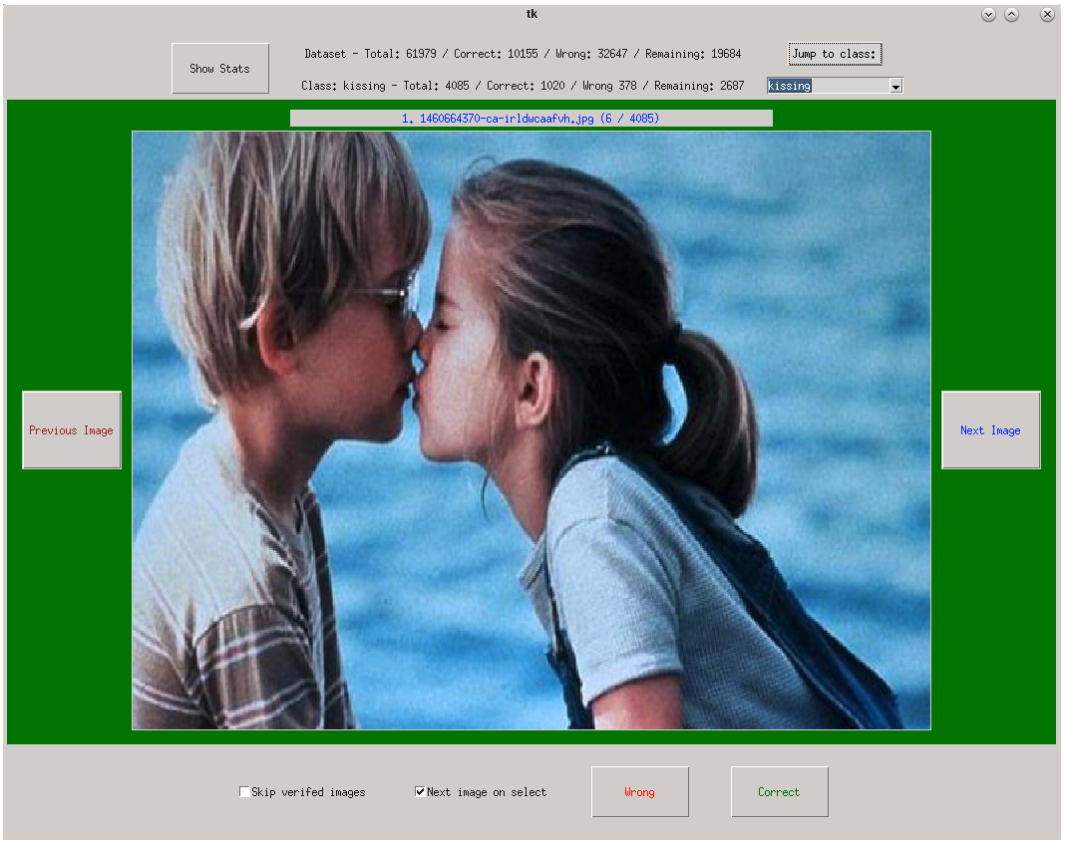


FIGURE 6.4: Screenshot from dataset verification application. Each image is checked by an human annotator and determined as correctly belongs to the corresponding class. In the example shown above, the class being verified is *Kissing*, and the image is determined as a correct instance of *Kissing* class.

fed to Google Images Download[1]. The results of these queries usually include too many false positives. In order to filter out these out-of-class images, we make use of the pose detection of MaskRCNN [1]. For each image, the poses are detected automatically by MaskRCNN, and the images containing less than two poses are discarded. Next, all remaining images are verified by human annotators, and wrongly labeled images are discarded. An application is created for the verification of images. Figure 6.4 is a screenshot from this application. The green background indicates that the image is already determined as a correct instance of *Kissing* class.

[1]https://github.com/hardikvasa/google-images-download

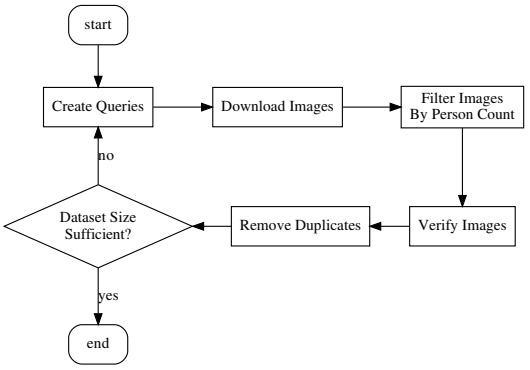


FIGURE 6.5: The process of collecting images. 100 images downloaded for each query by Google Images Download library. MaskRCNN[1, 16] is used to eliminate images containing less than 2 persons. Remaining images are verified visually and inappropriate images are excluded. Finally duplicate images are removed by the open source digiKam application.

Finally, duplicates are removed using the open-source digiKam[1] application. This process is repeated until the number of images in all classes is sufficient. As mentioned before, 1000 is selected as the minimum number of images for each class. Example images that are collected as a result of this process are given in Figure 6.3. Visualization of this visual data collection and cleaning process is given in Figure 6.5.

Another point to note is that since the images are collected from the web resources, most of them are real-world images; hence, an image may contain multiple interactions. For example, both kicking and punching may co-exist in a scene, or people may kiss at a party. In such cases, we look for the dominant interaction; if one interaction is solely dominant and can be clearly identified, we set it as the ground truth class; otherwise, the image is discarded.

At the end of this data collection process, the total amount of images in this extended dataset becomes 10502. The available set of classes is the same with the HII dataset, where the classes are *boxing-punching, dining, handshaking, high five, hugging, kicking, kissing, partying, speech* and *talking*, respectively. We call this new version of *HII* dataset as *Human Interaction Images v2 (HIIv2)*. The final dataset statistics are given in Table 6.1. Different from the HII dataset, HIIv2 is collected for the evaluation of our second work, which uses human poses as the primary feature set. Therefore, faces in HIIv2 are not annotated as in the HII dataset. HIIv2 is available online at `https://gtanisik.github.io/projects/hiiv2` address.

[1] www.digikam.org

7. EXPERIMENTAL RESULTS

7.1 Facial Descriptors

Implementation Details: The final face descriptor is formed by simply concatenating the proposed individual facial descriptors, which form a 31-dimensional feature vector. We then use Support Vector Machines (SVMs) with Gaussian-RBF kernel for learning separate classifiers for each descriptor type (face, scene, and deep). We evaluate the recognition performance in terms of average precision (AP) using five-fold cross-validation. During the training phase, we select a single kernel bandwidth and SVM cost parameter using a greedy search over the training set. We then use the same parameters in all folds, for all classes. After learning individual classifiers for each feature type, their responses are combined via late fusion. To this end, the prediction scores of individual classifiers are used as an input to a linear SVM. In this way, we aim to learn the optimal fusion weights of the descriptors. We use the output of this second layer SVM as the final classification score.

In our preliminary experiments, we experiment with different classifiers such as Random Forests, Neural Networks, and Bayesian classifiers. While we obtained comparable performances across these methods, we observed slightly better results with SVMs. Therefore, we use SVMs in all our experiments, which also simplifies the comparison of the proposed descriptors.

7.1.1 Results and Discussions

In this section, we first present an evaluation of the proposed facial descriptors in detail. Then, we evaluate the combinations of the facial descriptors together with GIST, BoW on dense SIFT, and CNN features.

TABLE 7.1: Average Precisions of the individual face descriptors using the face detection outputs

Feat.	b&p	din.	h.sh.	h.fv.	hug.	kick.	kiss.	pty.	sp.	tlk.	AVG
HFO	12.1	21.9	21.0	10.0	15.9	18.6	17.3	40.2	36.6	13.5	20.7
HFD	**15.7**	20.6	21.5	**15.7**	8.3	22.7	18.2	42.7	**37.8**	16.8	22.0
DF	15.6	24.6	23.3	12.1	14.3	5.9	23.7	23.4	21.8	13.2	17.8
CHFL	14.9	19.3	19.9	17.3	13.3	13.3	13.9	17.7	16.3	8.9	15.5
GHFL	14.8	25.2	22.3	12.4	**22.8**	**22.8**	18.2	27.0	18.1	14.1	19.8
Comb.	14.2	**30.0**	**24.8**	9.1	22.5	15.4	**48.0**	**44.4**	34.7	**17.9**	**26.1**

TABLE 7.2: Average Precisions of the individual face descriptors using ground truth faces

Feat.	b&p	din.	h.sh.	h.fv.	hug.	kick.	kiss.	pty.	sp.	tlk.	AVG
HFO	14.0	40.8	31.7	13.0	49.6	9.3	24.0	54.9	**91.8**	14.5	34.4
HFD	13.9	52.4	22.1	**16.1**	43.4	12.5	17.4	48.5	89.0	17.8	33.3
DF	15.8	41.8	28.7	9.3	38.9	**32.9**	51.3	36.6	65.0	**27.8**	34.8
CHFL	12.8	33.2	20.5	13.1	25.4	12.0	13.8	38.8	72.6	16.0	25.8
GHFL	10.7	48.1	21.9	6.2	32.4	17.4	8.2	46.0	70.1	14.2	27.5
Comb.	**17.9**	**57.1**	**41.2**	9.7	**61.8**	28.0	**85.3**	**69.4**	90.7	14.9	**47.6**

To evaluate the performance of facial descriptors, we have two settings: 1) Automatic: We use the output of the face detectors, 2) Face ground truth: We manually label each face region, together with the correct orientation of the face. Using the first setting, Table 7.1 presents the recognition rates of the proposed facial descriptors individually and combined. According to these results, Histogram of Face Directions (HFD) gives the best performance amongst the individual descriptors, whereas Circular Histogram of Face Locations (CHFL) performs the worst. Overall, the combination of all descriptors yields the best performance. We also observe that especially *highfive, boxing-punching* and *kicking* classes are difficult to identify. In these interactions, the relative spatial distributions of the faces tend to have a more random structure, and this affects the classification performance of facial descriptors.

By a qualitative evaluation of the dataset, we observe that the following conditions mainly affect the recognition performance: (1) In photographs, people usually look at the camera, so the natural orientations or directions of the faces can not be inferred. It is not easy to harvest images that depict the natural occurrences of the interactions. (2) Face detection algorithms may not find all the faces that are visible in the images. Although we use state-of-the-art face detection algorithms, many false negatives occur. (3) False positives also affect the performance of the facial descriptors negatively.

In order to evaluate the performance of the proposed descriptors independent of the errors introduced by the face detection, we annotated the faces in images with both location and orientation information and obtained the ground-truth faces. In this way, we aim to simulate the case with the perfect face detector. The results using this setting is presented in Table 7.2. The overall mean average precision of the combined descriptors with ground-truth faces rises from 26.1% to 47.6%. In other words, if we have perfect face detector and estimations for face orientations, we can achieve 47.6 mAP, even with relatively simple 31-dimensional face descriptors. There are also some interesting observations; An average precision as high as 91.8% is achievable using only HFO features for the *speech* class. This is not surprising,

TABLE 7.3: Comparison of the feature combinations using facial descriptors formed over the output of the face detectors.

Type	Feature	AP
Regular	FaceDesc	26.11
	GIST	43.52
	BoW	58.25
	BoW-SPM	58.68
Combined	BoW-SPM + GIST	59.46
	FaceDesc + GIST	47.13
	FaceDesc + BoW-SPM	59.77
	FaceDesc + GIST + BoW-SPM	60.60
Deep	Places-CNN	72.8
	Hybrid-CNN	77.86
Deep Combined	FaceDesc + Places-CNN	73.44
	FaceDesc + Hybrid-CNN	**78.33**
	GIST + BoW-SPM + Hybrid-CNN	75.05
	FaceDesc + GIST + BoW-SPM + Hybrid-CNN	76.09

TABLE 7.4: Comparison of the feature combinations using facial descriptors formed over the ground truth faces.

Feature	AP
GtFaceDesc	47.59
GtFaceDesc + GIST	56.18
GtFaceDesc + BoW-SPM	63.41
GtFaceDesc + GIST + BoW-SPM	64.48
GtFaceDesc + Places-CNN	76.05
GtFaceDesc + Hybrid-CNN	**80.11**
GtFaceDesc + GIST + BoW-SPM + Hybrid-CNN	79.56

as for the speech class, most of the faces are turned towards the speaker, which forms a discriminative case. In addition, for the *kissing* interaction, the combination of the facial features achieves a quite high mAP of 85.3%.

Next, we evaluate the performance of the scene features for interaction recognition. We first evaluate the individual performances of GIST, BoW, BoW with spatial pyramid matching (SPM) [44], and also the performances of deep features Places-CNN and Hybrid-CNN ([116]). After that, we evaluate the results for the combination of all feature sets and show the results in Table 7.3. We also report the combinations of scene features with facial descriptors formed using the ground truth face detections in Table 7.4.

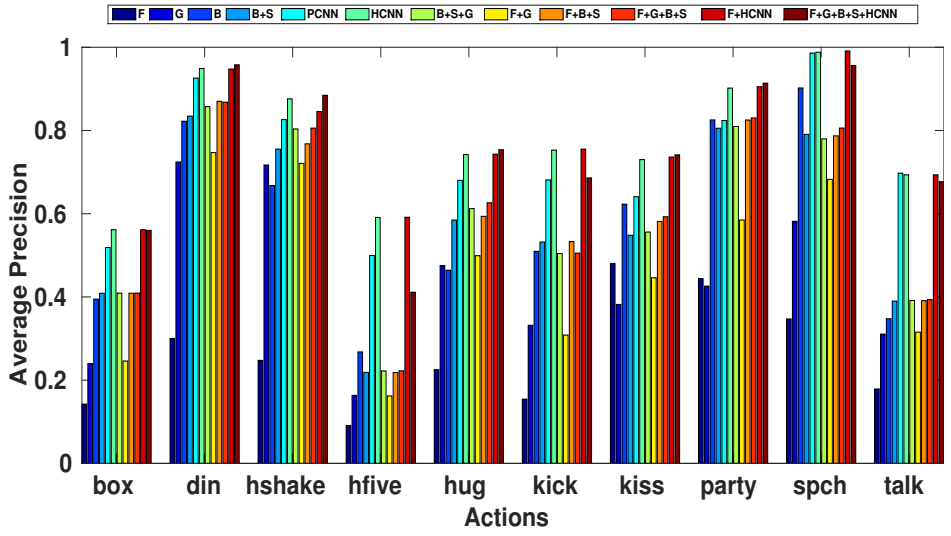


FIGURE 7.1: Average precisions per each of the interaction classes. In calculation of facial descriptions `FaceDesc`, the outputs of the face detectors are used. F:facial descriptors, G:GIST, B:BoW, B+S:BoW+SPM, PCNN:Places CNN, HCNN:Hybrid CNN.

Results in Table 7.3 and Table 7.4 show that, compared to the facial descriptors, scene features carry quite a lot of information for predicting the type of interaction. Proposed face descriptors contribute to the recognition of human interactions, and together with BoW and GIST features, they achieve a reasonable mAP of 60.6%, which would rise to 64.48% in the presence of a perfect face detector. On the other hand, the deep features are the most effective features for human interaction recognition, Places-CNN achieving 72.8% AP and Hybrid-CNN feature achieving an AP of 77.86%. An interesting observation at this point is that, when combining deep features with regular scene descriptors such as GIST and BoW, the performance slightly drops. This result suggests that GIST or BoW features do not carry any complementary information to the deep features. On the contrary, combining proposed facial descriptors with CNN features improves the performance, achieving 78.33% using the output of face detectors, and 80.11% when the ground truth face locations are used. We observe a similar pattern for both Places-CNN and Hybrid-CNN, while using Hybrid-CNN with facial descriptors yields better performance.

We present per class average precisions for our approach in Fig. 7.1. We observe that some of the interactions, such as boxing and high-five, are more difficult to recognize, having average precisions lower than 0.6. On the other hand, recognition performance on some interaction classes, such as dining and speech, is quite high. This is likely to be due to the similar spatial configurations present in these interactions. In all of the interaction classes, CNN features are very effective.

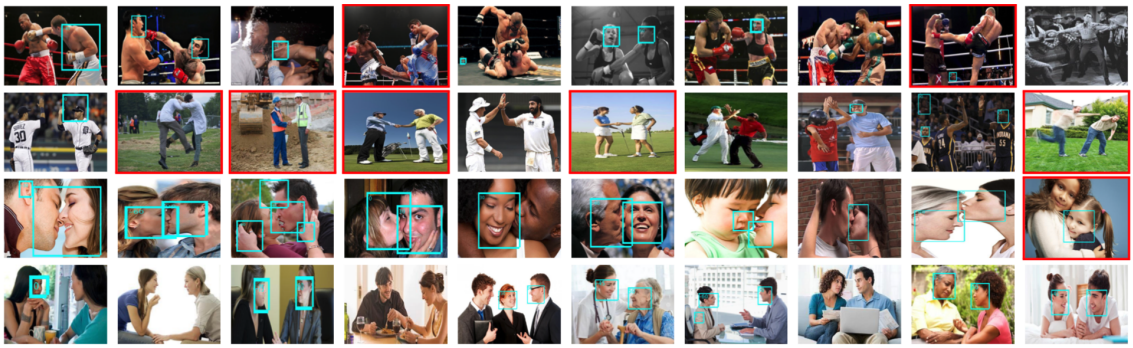


FIGURE 7.2: Images having the top ten scores using `FaceDesc` using automatic face detections, `GIST + BoW-SPM + Hybrid-CNN` features, for *boxing-punching, high-five, kissing* and *talking* interactions. Out-of-class images are shown with red borders and the face detection outputs (if any) are shown in cyan.

We also present the confusion matrix obtained using the best-performing feature combination, i.e., face descriptors and Hybrid CNN features, in Fig 7.3. This confusion matrix is achieved by assigning each test example the interaction class label that achieved the maximum prediction score. We observe that most of the confusion occurs between *boxing-punching* and *high five* interactions, and this is not surprising since these interactions have very similar spatial layouts with respect to interacting persons and their poses. *Kissing* and *hugging*, *handshaking* and *high five* are among the other interaction class pairs that are easier to be confused.

These results show that it is not easy to describe human interactions by looking at facial regions and their spatial layouts only (as demonstrated in Fig 1.1). Nevertheless, when used in combination with scene elements from the whole image, the facial descriptors can boost the recognition performance. CNN features are performing exceptionally well for this problem. However, we should note that CNN features are trained over large amounts of additional data. In addition, facial descriptors provide a quite compact representation, as stated in Table 7.5. Compared to the performance of random guessing (which is 10% in this dataset), we can say that facial descriptors provide valuable information about the human interactions and can be used complementarily to scene representations.

Qualitative results that are obtained as a result of using `FaceDesc + GIST + BoW-SPM + Hybrid-CNN` combination are presented in Fig. 7.2. As it can be seen, *high five* interaction is mostly confused with *handshaking* and *boxing-punching*, whereas *kissing* interaction is mostly confused with *hugging*.

TABLE 7.5: Feature dimensions of the utilized descriptors

Feature	FaceDesc	GIST	BoW	BoW+SPM	CNN
# of Dims	31	512	1000	4000	4096

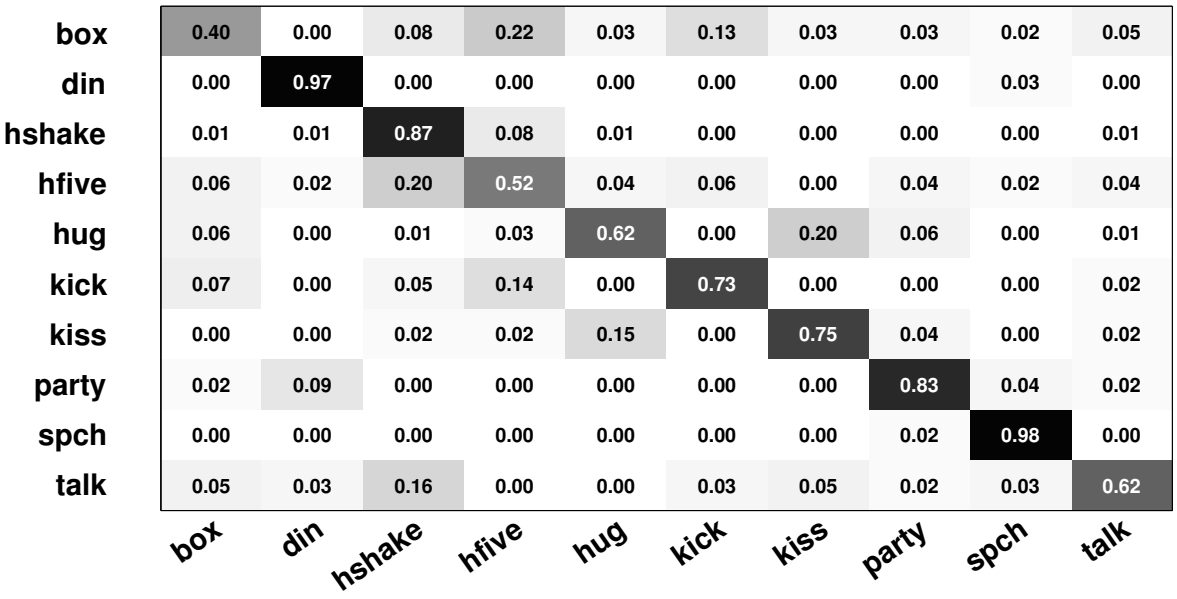


FIGURE 7.3: Confusion matrix for interaction classification using FaceDesc and Hybrid CNN features in combination.

One of the essential characteristics of the proposed facial descriptors is their efficiency. Table 7.5 shows the dimensionalities of each of the utilized descriptors. Our proposed 31-dimensional face descriptors provide 2-6% points increase in average precision when used with BoW+SPM and/or GIST, and 1-3% points increase when used with the CNN features, without requiring excessive training.

7.2 Multi-stream Pose CNNs

7.2.1 Implementation Details

To evaluate the effectiveness of the proposed multi-stream pose networks, individual CNN streams are trained by fine-tuning the CNN architectures AlexNet [3], DenseNet [5], ResNet [4], Inception-V3 [118], and VGG-19 [119] that were pretrained on ImageNet [45]. ResNet [4] is used as the major architecture to test, tune and make inferences since it is faster to train and performs generally better than DenseNet, Inception, and VGG architectures. On the other hand, to make a fair comparison with the previous work [120], results for AlexNet is reported as well.

The hyper-parameters for the proposed models are optimized via grid search. Since there are two datasets (HII and HIIv2), five different CNN backbones, four single-stream and seven

TABLE 7.6: Accuracies of single and multiple streams using the base ResNet152 architecture on the HII dataset. Best accuracies are shown in red, and the second-bests are shown in boldface.

Model	b&p.	din.	h.sh.	h.fv.	hug.	kick.	kiss.	pty.	sp.	tlk.	overall accuracy
RGB	79.67	95.04	90.91	**87.62**	83.69	89.16	90.20	87.04	95.65	89.29	88.85
Pose Mask (PM)	53.57	91.55	85.71	75.73	64.12	79.07	67.83	70.83	89.60	78.33	76.35
Pose Only (PO)	67.83	90.65	91.72	79.61	82.01	84.71	86.00	84.31	94.31	85.95	85.14
Pose Highlighted (PH)	72.31	94.20	88.74	82.00	85.93	90.24	91.09	85.98	94.12	85.95	86.99
RGB + PM	**81.03**	95.77	**94.04**	86.00	86.76	**94.25**	89.72	84.00	95.08	89.43	89.86
RGB + PO	76.42	**96.45**	93.33	85.98	85.93	90.70	90.20	**88.89**	**96.61**	91.23	89.70
RGB + PH	79.39	**96.45**	91.78	84.31	**89.39**	**93.02**	**93.20**	85.19	95.73	**94.92**	**90.37**
RGB + PM + PO	**80.95**	**96.40**	**93.42**	84.85	**88.24**	91.95	92.16	87.62	94.92	**91.67**	**90.37**
RGB + PM + PH	80.33	95.71	93.33	**88.24**	87.59	92.13	**92.31**	87.38	**95.80**	91.53	**90.54**
RGB + PO + PH	77.86	96.40	92.72	83.67	88.06	91.76	91.09	**89.52**	95.80	90.91	89.86
RGB + PM + PO + PH	**80.95**	94.96	92.72	84.21	84.51	92.13	89.80	87.62	95.73	88.52	89.19

multi-stream model combinations (see Table 7.8), it is impractical to run a grid-search for every combination. Three methods are used in order to create a reasonable search space. 1) Only the HIIv2 dataset is used, and the same parameter sets are used for HII. 2) Parameters are optimized on `Pose Highlighted` for single-stream models and on `RGB + Pose Highlighted` multi-stream models. 3) Initially, a wide range of parameter set is used to optimize the ResNet152 model. Afterward, only the most effective values are used for other network backbones. As a result, five network backbones are optimized on the HIIv2 dataset once for single-stream and once for multi-stream representations. By pruning the ineffective options on the single-stream ResNet152 model, the number of parameter sets to be searched is reduced to 24 for each network model and stream combination.

In all evaluations, the accuracy metric is reported, which is defined as the ratio of correct classifications vs. all classifications. In addition, mean average precision (mAP) is used to compare the proposed approach with related works.

7.2.2 Experimental Results

We have run a series of analyses to assess the performance of the proposed pose-based streams. In the first set of our experiments, we evaluate the classification accuracies of every single stream and the possible fused architectures.

7.2.2.1 Evaluations of single streams

The classification accuracies of each single-stream model that are using the fine-tuned ResNet152 [4] architecture are given in Table 7.6 and Table 7.7 for the *HII* and *HIIv2* datasets

respectively. The first row, labeled as RGB, represents the base model that operates on the RGB input. As can be observed from these results, amongst the single streams that operate over RGB, PO, PM, and PH inputs individually, the RGB stream achieves the best results with the accuracies of 88.85% and 91.35%. From these results, one can say that human-human interactions are greatly affected by the images' overall context. Amongst pose based streams, the `PH` stream gives the highest accuracy,

An interesting point to note is the relatively high performance of the PM stream on the classification of the interactions. In PM stream, the input is the binary masks extracted using the person segmentation of MaskRCNN. As can be seen from the example images, it is even quite hard for the human eye to differentiate the interaction classes by looking at the pose masks alone. In this respect, gaining a classification accuracy of $\approx 77\%$ is quite impressive. This strongly supports the hypothesis that poses can be regarded as strong indicators of the interaction classes. Another inference that can be made by the performance of the Pose Mask (PM) Model is that some interactions could be defined solely by the composition of poses. *Dining*, and *handshaking* could be given as an example to these interactions. PM stream yields lower classification accuracies on the rest of the classes because each of those has similar characteristics with one or more of other classes. For example, *boxing-punching* and *high-five* are similar because they both are defined by the composition of hands. *Partying* and *speech* usually performed by many people, and the context is needed to determine actual interaction.

The plots of ResNet152 model loss and accuracy on both train and validation sets are given in figures 7.4, 7.5, 7.6, 7.7 respectively for single-stream `RGB, Pose Masks, Pose Only,` and `Pose Highlighted` inputs.

7.2.2.2 Evaluation of multiple streams

Multi-stream network versions are evaluated based on the results of early-fusion. Besides, a comparison between early and late fusion is given below.

Three different multi-stream network variants are tested, as defined in Section 3.4.1. (1) Two-stream networks are created by combining Pose Masks(PM), Pose Only(PO) and Pose Highlighted(PH) with RGB, which are referred respectively as `RGB+PM`, `RGB+PO` and `RGB+PH`. (2) Three-stream networks are created by RGB and two-way combinations of pose networks,

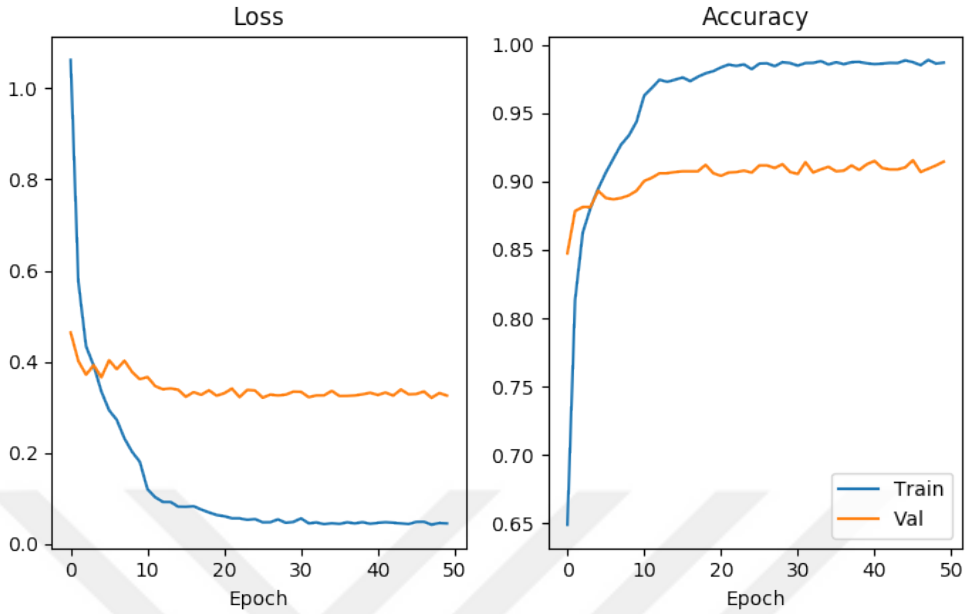


FIGURE 7.4: Plot of ResNet152 model Accuracy and Loss on both train and validation sets for the `RGB` stream.

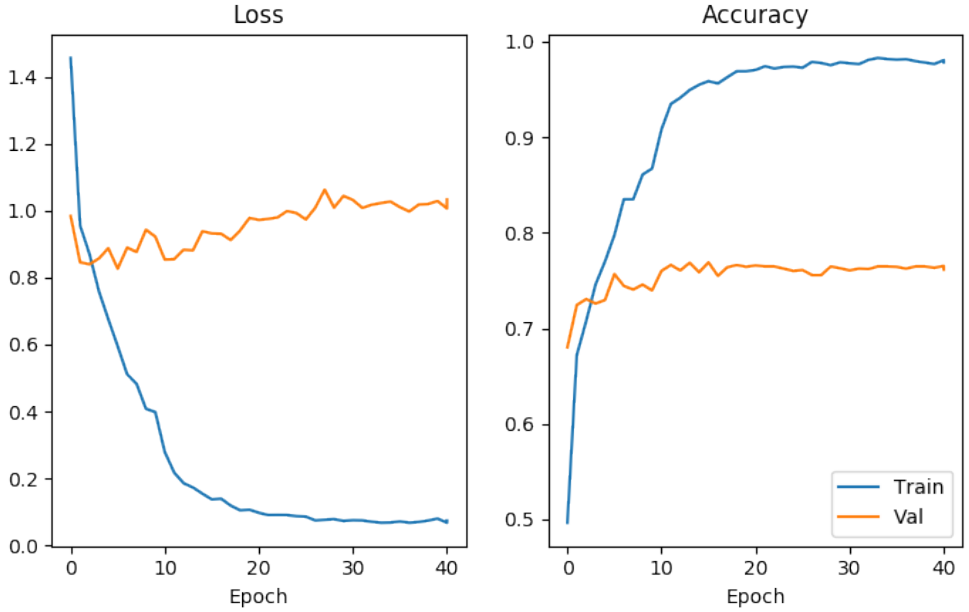


FIGURE 7.5: Plot of ResNet152 model Accuracy and Loss on both train and validation sets for the `Pose Masks` stream.

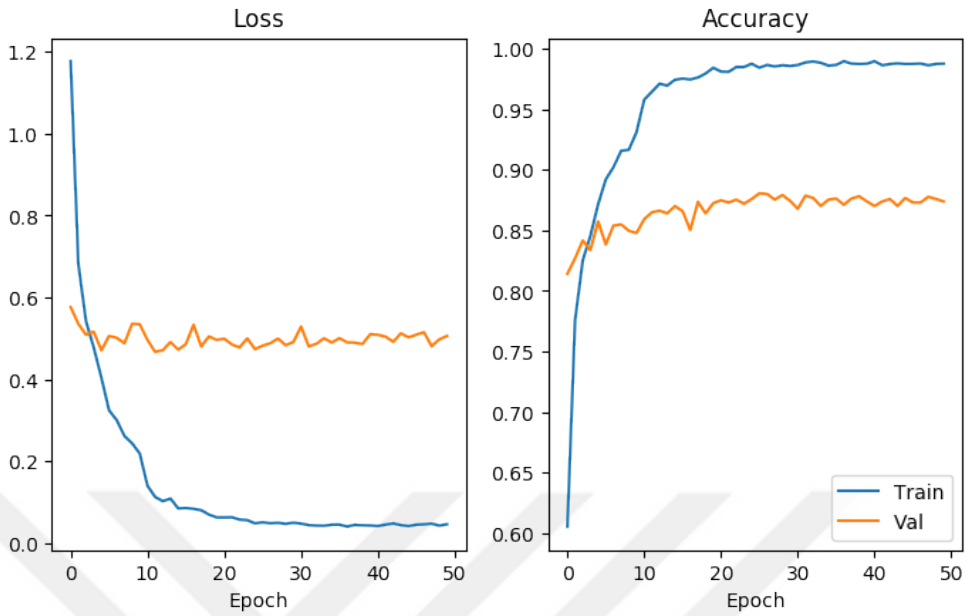


FIGURE 7.6: Plot of ResNet152 model Accuracy and Loss on both train and validation sets for the `Pose Only` stream.

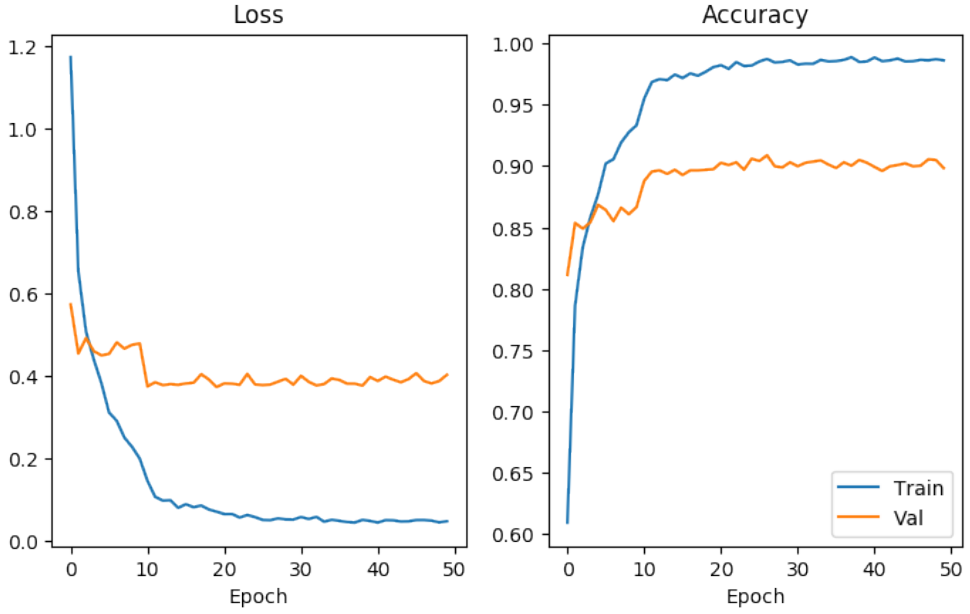


FIGURE 7.7: Plot of ResNet152 model Accuracy and Loss on both train and validation sets for the `Pose Highlighted` stream.

TABLE 7.7: Accuracies of single and multiple streams using the base ResNet152 architecture on HIIv2 dataset. Best accuracies are shown in red, and the second-bests are shown in boldface.

Model	b&p.	din.	h.sh.	h.fv.	hug.	kick.	kiss.	pty.	sp.	tlk.	overall Accuracy
RGB	84.00	96.30	95.00	88.73	90.50	93.97	**96.15**	90.95	93.02	83.50	91.35
Pose Mask (PM)	56.50	90.28	90.42	74.02	65.00	87.44	85.10	69.05	82.33	66.00	77.01
Pose Only (PO)	73.50	93.52	96.25	83.33	89.50	92.46	92.31	84.76	88.37	81.00	87.72
Pose Highlighted (PH)	79.00	94.91	97.08	87.25	89.00	91.96	**95.67**	89.52	91.63	85.50	90.34
RGB + PM	**87.50**	**98.15**	**97.92**	83.82	89.50	95.48	**96.15**	92.38	93.95	88.00	92.45
RGB + PO	84.50	96.30	**97.92**	87.25	91.00	93.47	**96.15**	90.95	**95.35**	85.50	92.02
RGB + PH	84.50	97.22	**97.50**	**90.20**	**92.50**	95.98	**96.15**	**92.86**	**95.81**	85.50	**92.97**
RGB + PM + PO	83.00	96.30	97.08	89.71	87.00	94.97	**96.15**	90.95	93.95	88.00	91.87
RGB + PM + PH	83.00	**97.69**	96.67	**90.69**	**94.50**	**95.98**	**96.15**	91.90	94.42	**90.50**	**93.26**
RGB + PO + PH	**86.50**	96.76	96.25	88.24	**92.50**	**96.48**	**96.15**	91.90	**95.81**	85.00	92.69
RGB + PM + PO + PH	**87.50**	95.37	95.83	84.80	91.00	93.47	**96.15**	**92.86**	93.95	**93.00**	92.50

TABLE 7.8: Comparison of different base CNN architectures. The best accuracies for each network type (row) are shown in red, and the second-bests are shown in boldface. The model with the best accuracy on a specific network (column) is shown as underlined.

	HII dataset					HIIv2 dataset				
Model	AlexNet	DenseNet	ResNet	Inception V3	VGG 19	AlexNet	DenseNet	ResNet	Inception V3	VGG 19
RGB	76.52	87.16	**88.85**	**90.30**	88.01	79.49	**91.40**	**91.35**	90.30	88.86
Pose Mask (PM)	67.40	**78.89**	76.35	77.01	**77.70**	73.37	**79.25**	77.01	77.01	**77.44**
Pose Only (PO)	71.11	84.97	**85.14**	**86.42**	82.94	80.16	**88.05**	**87.72**	86.42	86.76
Pose Highlighted (PH)	67.57	**86.99**	**86.99**	**89.34**	84.46	77.63	**89.96**	**90.34**	89.34	88.67
RGB + PM	81.25	**90.37**	89.86	**90.58**	88.34	78.82	**91.40**	**92.45**	90.58	90.39
RGB + PO	79.73	87.67	**89.70**	**91.30**	87.84	78.30	**92.26**	92.02	91.30	**90.20**
RGB + PH	78.89	89.53	**90.37**	**91.35**	87.50	78.20	**92.35**	**92.97**	91.35	89.53
RGB + PM + PO	82.43	90.20	**90.37**	**90.54**	87.84	82.36	**91.16**	**91.87**	90.54	89.82
RGB + PM + PH	81.59	89.70	**90.54**	**91.35**	87.50	82.60	**91.97**	**93.26**	91.35	89.82
RGB + PO + PH	80.57	88.01	**89.86**	**91.01**	87.16	80.35	**92.11**	**92.69**	91.01	90.20
RGB + PM + PO + PH	82.09	88.85	**89.19**	**91.54**	87.33	80.50	**91.92**	**92.50**	91.54	89.53

referred as RGB+PM+PO, RGB+PM+PH and RGB+PO+PH. (3) The last architecture is formed by combining all the streams together, which is referred to as RGB+PM+PO+PH.

At first glance, it can be observed that almost all of the multi-stream models are superior to the single RGB stream. This suggests that giving more attention to poses increases the performance of the human-human interaction recognition task. In general, the three-stream models are superior to two-stream ones. RGB + PM + PH model achieves the best performance among all approaches on both datasets. On the other hand, four-stream model performs slightly worse than RGB + PM + PH. The main reason for this can be the raised ambiguity in the PM masks, in addition to the false-negative pose detections or the lack of some limbs (such as a fist missing in a boxing-punching scene). Another issue that the proposed models suffer from is that images may contain multiple interactions. When the pose detections are not complete, the model can be biased to secondary interactions in the image.

Table 7.8 shows the comparative results of the proposed approach using different base CNN architectures such as AlexNet, DenseNet, ResNet , Inception-V3 and VGG-19 on each dataset. Pre-trained versions of these architectures provided by PyTorch are used to test the proposed approach. For all architectures, the best performing versions are chosen which are ResNet152, DesNet161, and VGG19. Since Inception-V4 [121] is not officially provided by PyTorch yet, Inception-V3 is used in benchmarks. From Table 7.8, one can observe that deep models have benefited from the additional amount of data available in the extended version *HIIv2* since the models trained on this dataset performs better than models trained on the previous version *HII* for almost all cases. This confirms that when the over-fitting problem is solved, there is more to learn about recognizing interactions using deep architectures. Overall, ResNet performs best on HIIv2 and Inception-V3 on HII. Since ResNet produces better results overall, the proposed models are based on ResNet in the rest of the experiments.

Confusion matrix obtained on *HIIv2* dataset using the best performing architecture `RGB + PM + PH` is presented in Figure 7.8. Rows of the matrix are actual ground-truth labels and columns are predictions. Highest confusion is between *boxing-punching* and *high-five* classes. This is expected because spatial layouts of people in these interactions are similar and they are both characterized by the position of arms. The rest of the confusion looks well distributed with small reasonable peaks. For example *kissing* and *hugging* classes have a big potential to be confused. Example images can be seen on Figure 7.9 (fourth row).

Some successful and failed predictions of the proposed approach are given in Figure 7.9 and Figure 7.10. As can be seen, some classes such as *kissing* and *hugging* or *handshaking* and *high five* may look similar. Another point that could be inferred by the figure is the second misclassified image of *boxing-punching* (first two rows, fourth column). In the original image, one can see that the man in the middle tries to punch the man on the right. The leftmost person is trying to separate them.

On the other hand, in the Pose Highlighted version of the scene, the man on the right is faded out because it is not detected. Thus, the Pose Highlighted version sees two persons hugging. This is an example, where there can be multiple interactions in the same image, and our model is able to find one of them.

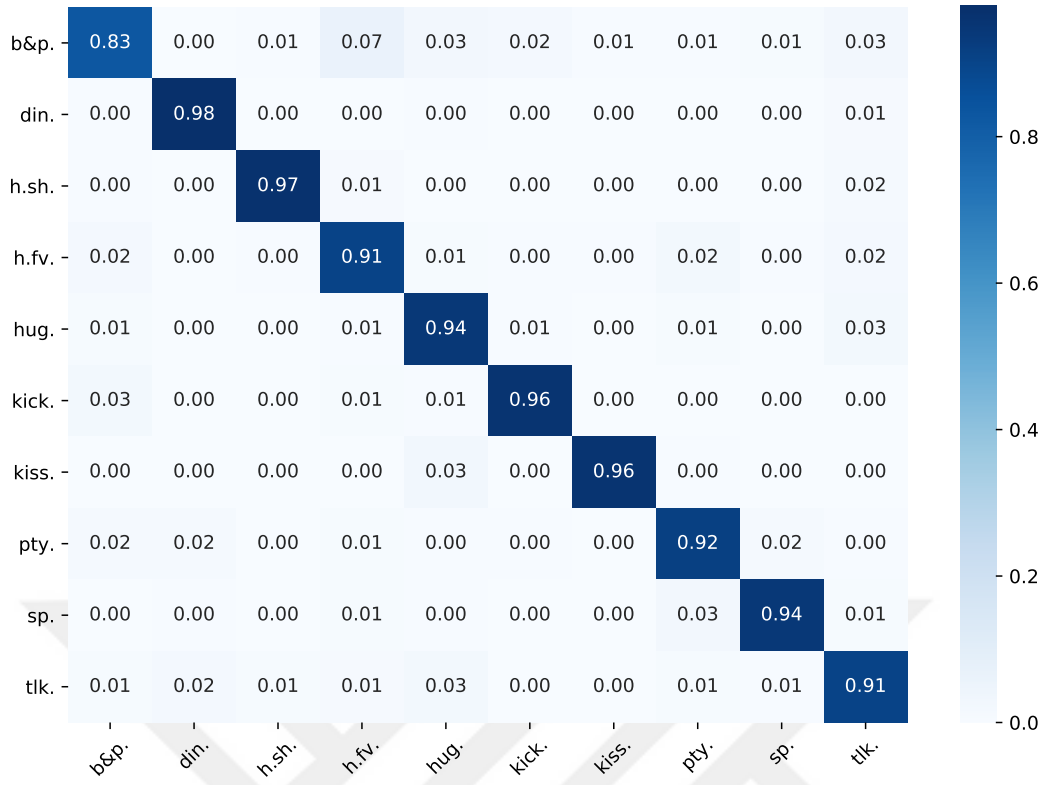


FIGURE 7.8: Confusion matrix for interaction classification using `RGB + Pose Masks + Pose Highlighted` model on *HIIv2* dataset. Rows are actual labels and columns are predictions.

The plots of ResNet152 model loss and accuracy on both train and validation sets are given in figures 7.11, 7.12, 7.7, respectively for two-stream `RGB+PM`, `RGB+PO` and `RGB+PH` inputs. Figures 7.14, 7.16, and 7.16 represents the three-stream `RGB+PM+PO`, `RGB+PM+PH`, and `RGB+PO+PH`. Finally, 7.17 represents the four-stream `RGB+PM+PO+PH` version.

7.2.2.3 Fusion point selection for early fusion

Experimental results on selecting the fusion point for multi-stream networks are given in Table 7.9. As mentioned earlier in Section 5.4.3, ResNet has four main blocks and a final classification layer at the end.

To find the best point to fuse the streams, five different fusion schemes are created. Each of these fusion schemes is tested with proposed multi-stream networks. Only the best performing two-stream and three-stream models are given in the Table 7.9 for clarity. The first four columns of the table show the training accuracy of networks that are fused right before the corresponding block. For example, the first column of the table indicates the multi-stream

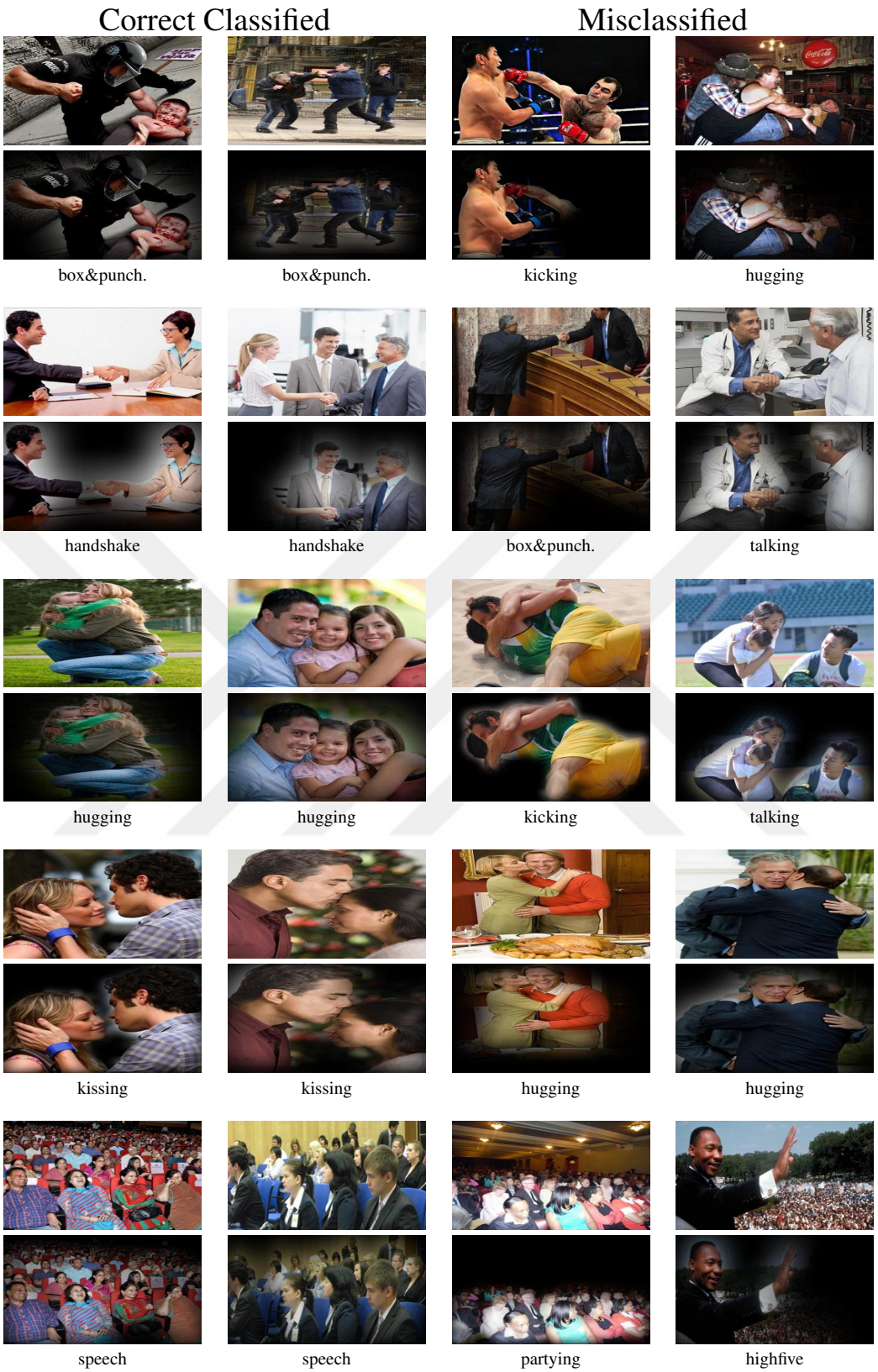


FIGURE 7.9: Sample classification results of `RGB + Pose Mask + Pose Highlighted` on *boxing-punching, handshaking, hugging, kissing* and *speech* classes. Each row contains correct and wrong classified images for one class. Two rows are shown for each class: (1) original image and (2) our Pose Highlighted image corresponding to the original one. Misclassified samples are shown with their predicted categories. Rows sequentially represent *boxing-punching, handshaking, hugging, kissing* and *speech*. This figure is better examined by zooming in digitally.

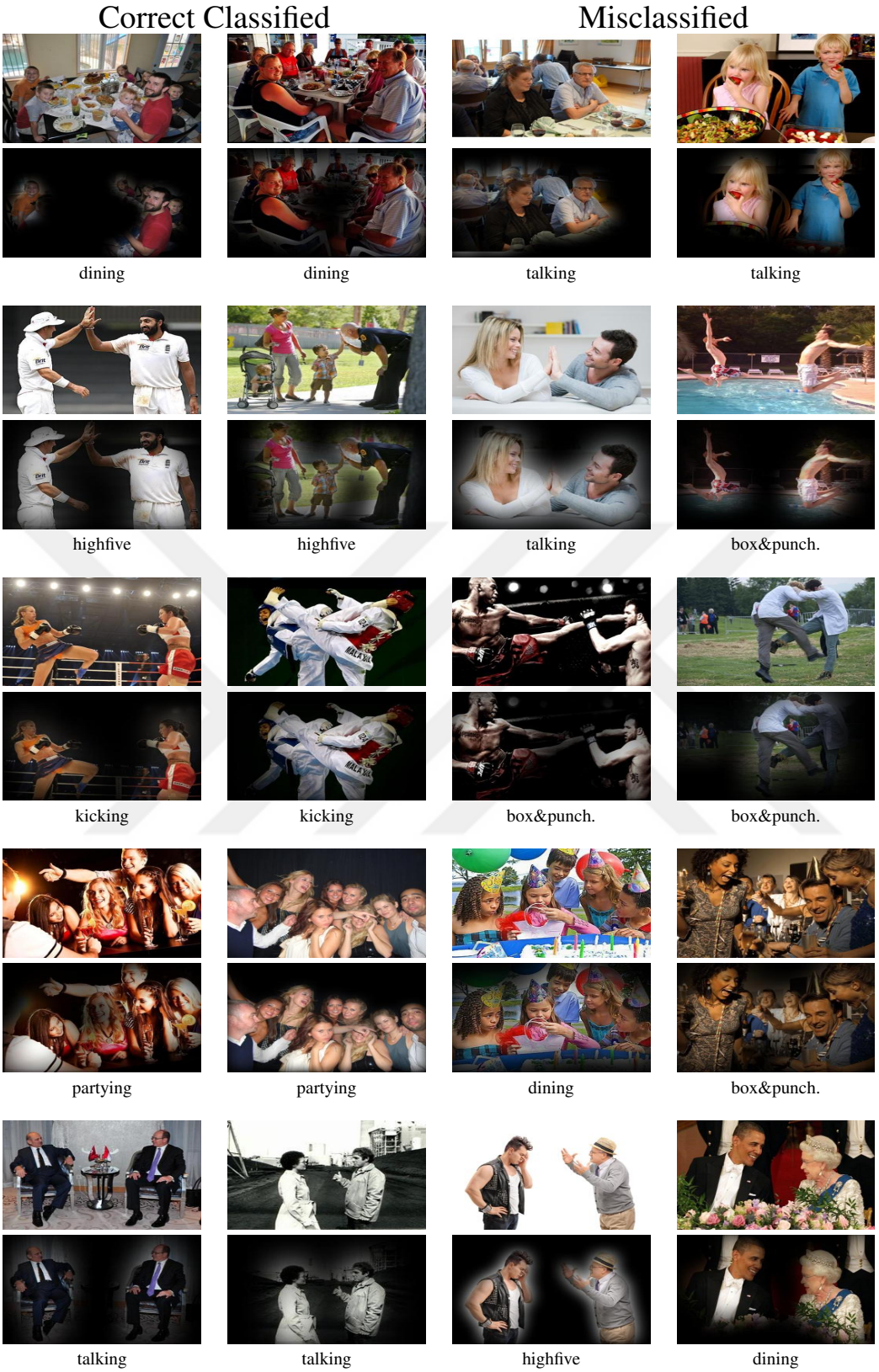


FIGURE 7.10: Sample classification results of `RGB + Pose Mask + Pose Highlighted` on *dining, highfive, kicking, partying* and *talking* classes. Each row contains correct and wrong classified images for one class. Two rows are shown for each class: (1) original image and (2) our Pose Highlighted image corresponding to the original one. Misclassified samples are shown with their predicted categories. Rows sequentially represent *dining, highfive, kicking, partying* and *talking*. This figure is better examined by zooming in digitally.

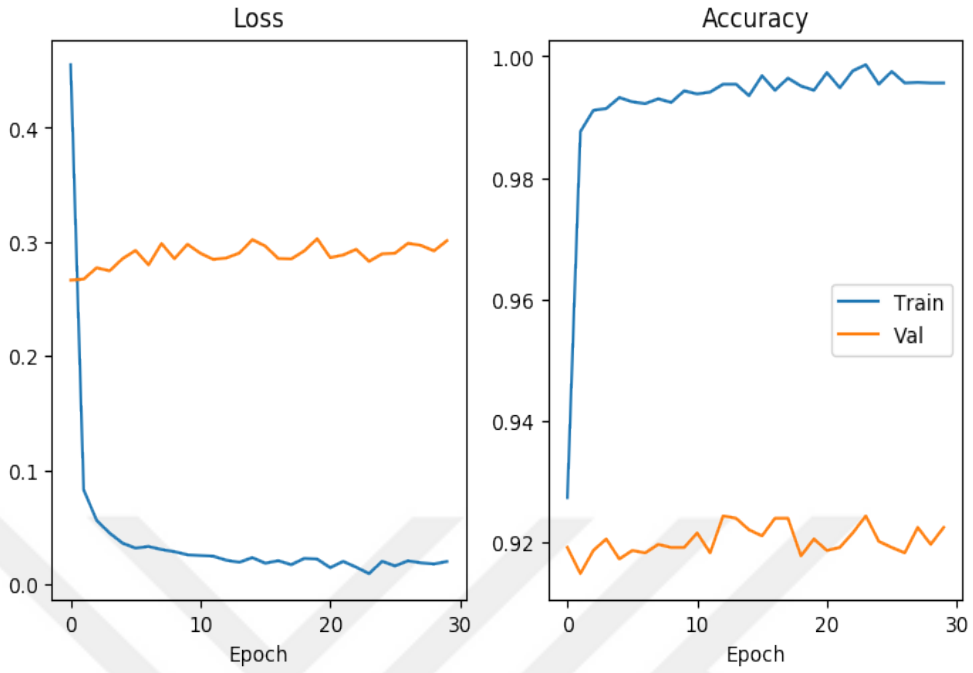


FIGURE 7.11: Plot of ResNet152 model Accuracy and Loss on both train and validation sets for the `RGB + Pose Masks` stream.

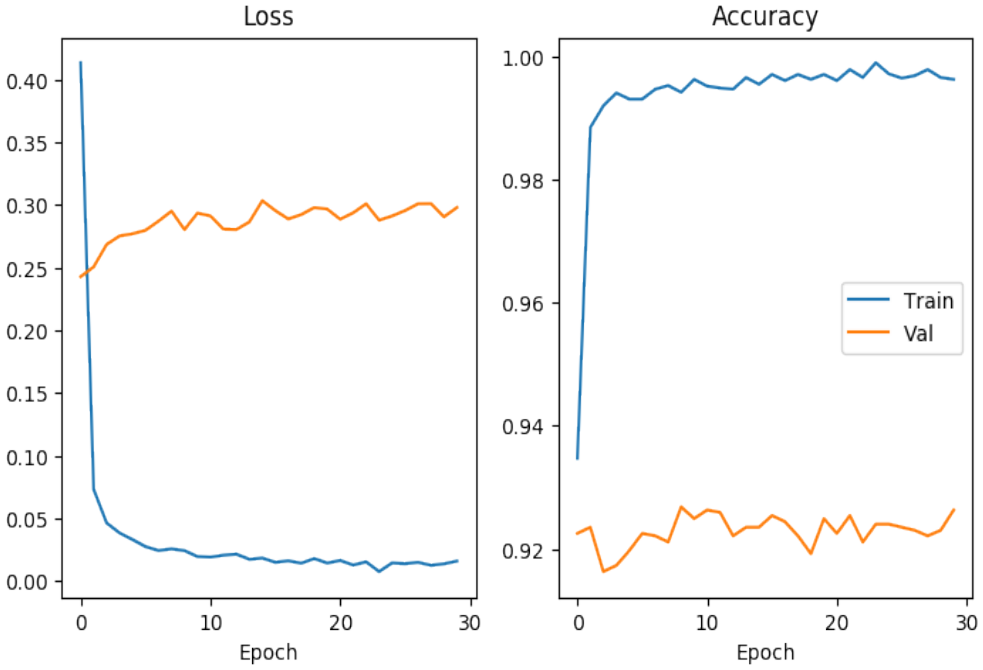


FIGURE 7.12: Plot of ResNet152 model Accuracy and Loss on both train and validation sets for the `RGB + Pose Only` stream.

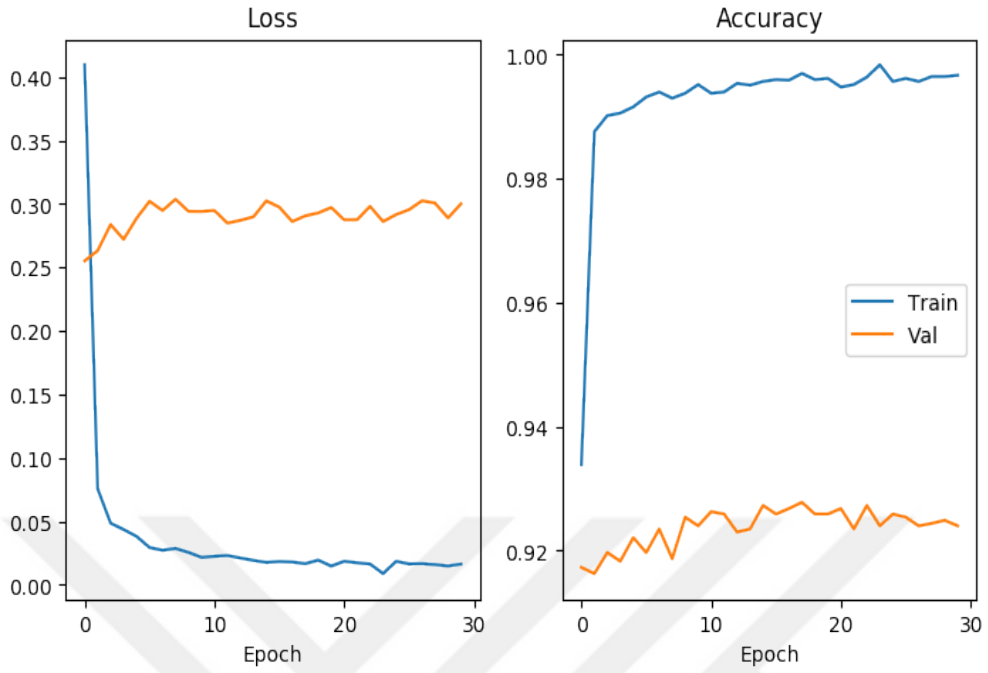


FIGURE 7.13: Plot of ResNet152 model Accuracy and Loss on both train and validation sets for the `RGB + Pose Highlighted` stream.

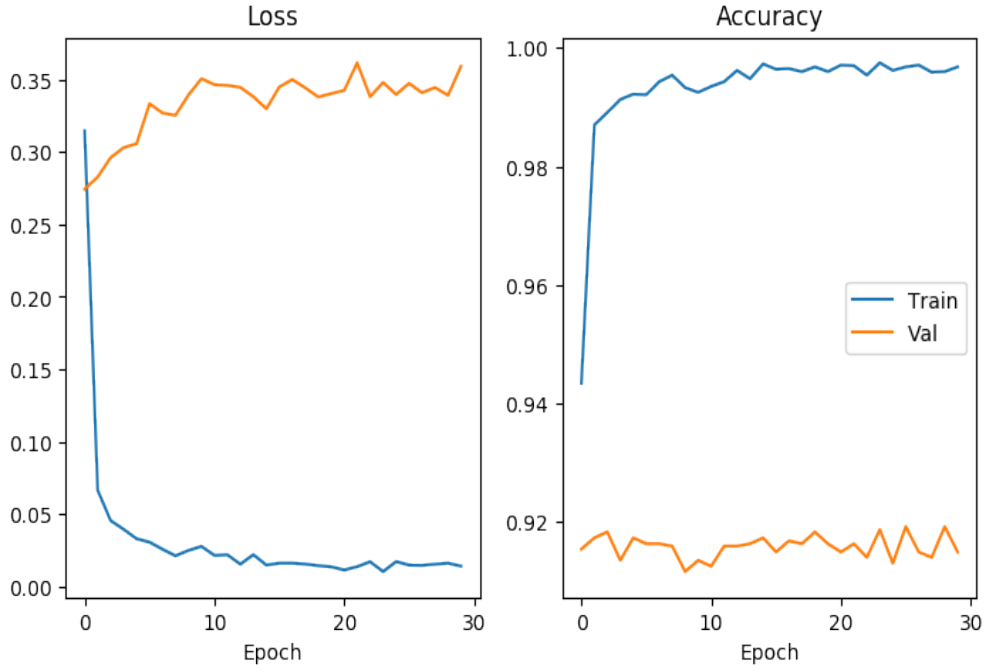


FIGURE 7.14: Plot of ResNet152 model Accuracy and Loss on both train and validation sets for the `RGB + Pose Masks + Pose Only` stream.

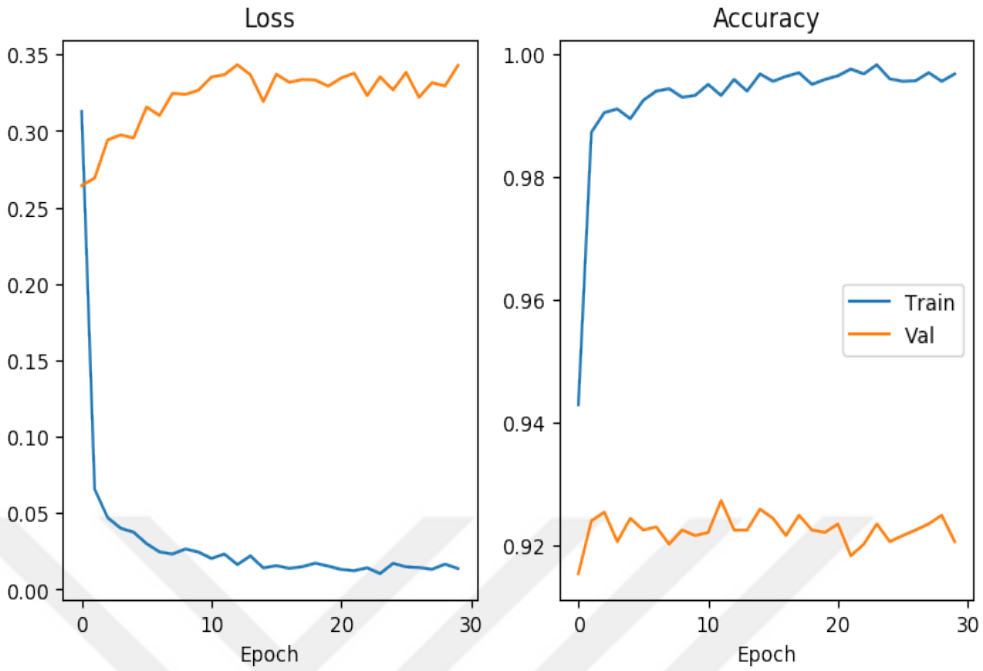


FIGURE 7.15: Plot of ResNet152 model Accuracy and Loss on both train and validation sets for the `RGB + Pose Masks + Pose Highlighted` stream.

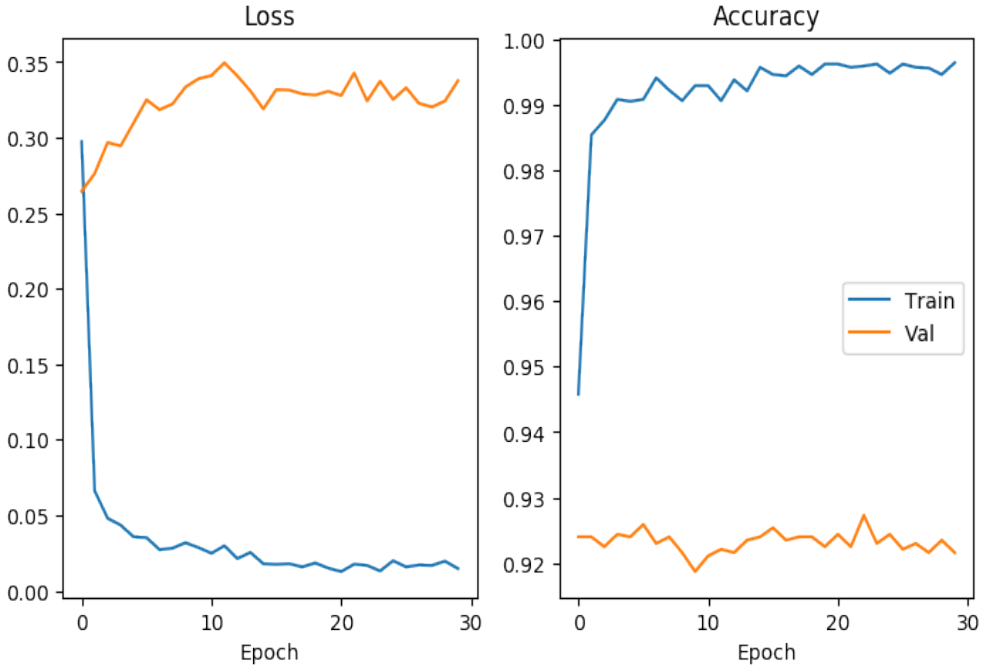


FIGURE 7.16: Plot of ResNet152 model Accuracy and Loss on both train and validation sets for the `RGB + Pose Only + Pose Highlighted` stream.

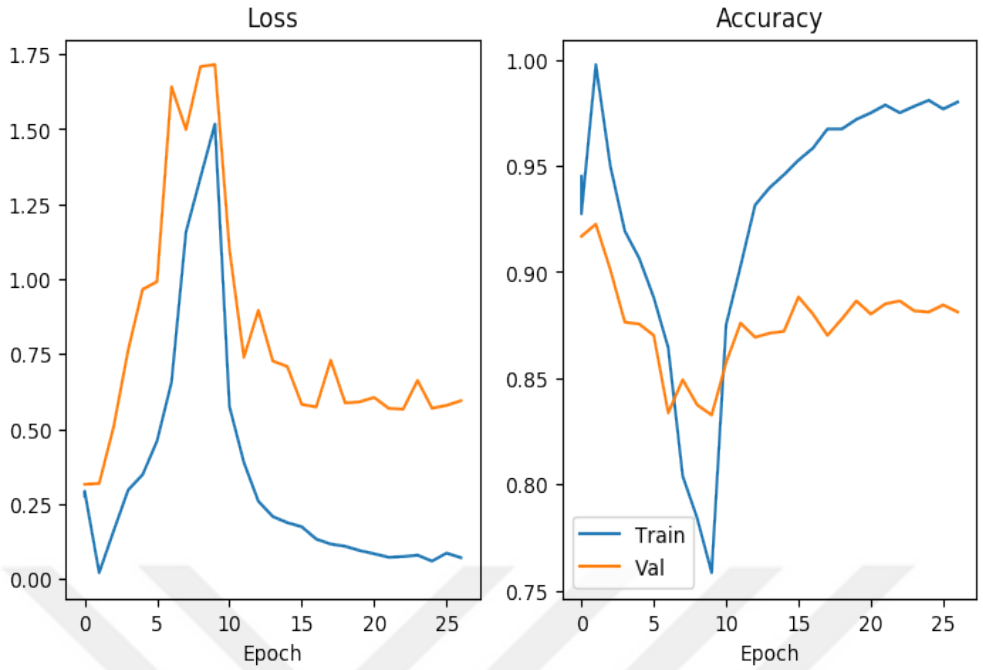


FIGURE 7.17: Plot of ResNet152 model Accuracy and Loss on both train and validation sets for the `RGB + Pose Masks + Pose Only + Pose Highlighted` stream.

TABLE 7.9: Effect of fusion point location on ResNet-based multi-streams over HIIv2 dataset.

	Before block 1	Before block 2	Before block 3	Before block 4	Before softmax
`RGB + PH`	91.16	91.30	90.92	91.73	**92.97**
`RGB + PM + PH`	90.73	91.92	90.49	91.68	**93.26**

network in which the streams are fused right before block 1. Similarly, the last column indicates the multi-stream network in which the streams are fused before the classification layer. From these results, one can say that the success of the network increases as the fusion point gets closer to the end.

7.2.2.4 Early vs Late Fusion

The main difference between early and late fusion is that early fusion is trained with an end-to-end fashion, which means that the final model is trained by looking at all inputs from every stream. Late fusion, on the other hand, uses the scores of individually trained models for each stream. Hence, it does not require an additional training phase.

A comparison between early and late fusion methods is given in Table 7.10. It shows that

TABLE 7.10: Comparision of Early and Late Fusion

	Early Fusion		Late Fusion	
	Accuracy	mAP	Accuracy	mAP
RGB + PM	**92.45**	**95.82**	91.83	94.81
RGB + PO	92.02	**96.06**	**92.30**	95.85
RGB + PH	**92.97**	**96.49**	92.69	95.99
RGB + PM + PO	91.87	**95.66**	**92.30**	95.52
RGB + PM + PH	**93.26**	**96.09**	93.21	95.98
RGB + PO + PH	92.69	**96.18**	**92.88**	96.16
RGB + PM + PO + PH	92.50	**96.21**	**93.21**	96.08

late fusion is a powerful alternative and has a performance close to early fusion. Better mAP performance of Early Fusion shows that it has more confidence when making inferences. Another important interpretation would be that late fusion has better accuracy when the number of streams increases. However, this result would be specific to our case, and this inference requires many more tests to be certain.

7.2.2.5 Significance analysis of the contribution of proposed pose-based inputs

We have shown that the proposed pose-based inputs are capable of representing interactions individually. They also provide a valuable contribution when combined with the original RGB image. To check if this contribution observed by chance or it is the effect of pose-based inputs, we have conducted McNemar's Statistical Test [122] against baseline RGB stream and proposed multi-stream models. The test is used to see the effects of a method compared to a baseline. It requires dichotomous scores and nominal or ordinal scale measurements to be applied. A null hypothesis is tested that unlike pairs of responses (0, 1) and (1, 0) are equally likely. In the domain of a classification problem, (0, 1) corresponds to samples that the first classifier fails, but the second one succeeds to classify. Similarly (1, 0) corresponds to samples that the first classifier succeeds, but the second one fails to classify. A 2x2 contingency table is created to apply the test. Table 7.11 shows the contingency table to apply significance test for our proposed `RGB + PM + PH` model. It's compared to vanilla single-stream `RGB` model to see if there is a significant contribution of `PM + PH` streams to `RGB` stream. The (0, 1) pairs correspond to (negative, positive) cell, and (1, 0) pairs correspond to the (positive, negative) cell in the table. The test statistic is given by the formula:

$$\chi^2 = \frac{[\#(0,1) - \#(1,0)]^2}{\#(0,1) + \#(1,0)} \tag{16}$$

that is in our case $(74-36)^2/(74+36) = 13.127$. The corresponding probability (*p-value*) can be looked up on the χ^2 table or calculated by a software that is 0.0. To reject a null

TABLE 7.11: Contingency table to apply McNemar's test between Original RGB stream and multi-stream RGB +PM + PH model

	RGB + PM + PH positive	RGB + PM + PH negative
RGB positive	1877	**36**
RGB negative	**74**	105

TABLE 7.12: Statistical Significance Test Results

RGB baseline vs Proposed multi-stream architectures	p-value	Accuracy
RGB	-	91.35
RGB + Pose Masks(PM)	0.040	92.45
RGB + Pose Only(PO)	0.239	92.02
RGB + Pose Highlighted(PH)	**0.000**	92.97
RGB + PM + PO	0.435	91.87
RGB + PM + PH	**0.000**	93.26
RGB + PO + PH	0.009	92.69
RGB + PM + PO + PH	0.054	92.50

hypothesis test that claims there is no statistical significance between two methods, the *p-value* is supposed to be close to 0. It is common sense to select a cut-off value, which is generally used as 0.005. In our case, the *p-value* is 0.0; therefore, McNemar's test suggests that it is almost impossible that the success of RGB+PM+PH stream is observed by chance.

Table 7.12 shows the test results for trained models on HIIv2 dataset. Four of the models has a *p-value* that is less than the cut-off value, and one very close to it. This suggests that the contribution of proposed multi-stream models is very unlikely to happen by chance.

7.2.2.6 Visualization of network activations

Visualization of the network layer helps us understand what is actually learned by the architecture. Early studies [123, 124] tries to visualize the weights of convolution and FC filters. Meanwhile, recent papers [125, 126] focus on the salient regions, which has the most effect on making a decision. In this work, we use a very recent paper called Score-CAM [126] to visualize the salient regions, which affects most when predicting the interaction in the image. One figure is given for each class. For the sake of simplicity RGB, PH and RGB + PH streams are presented in the figures.

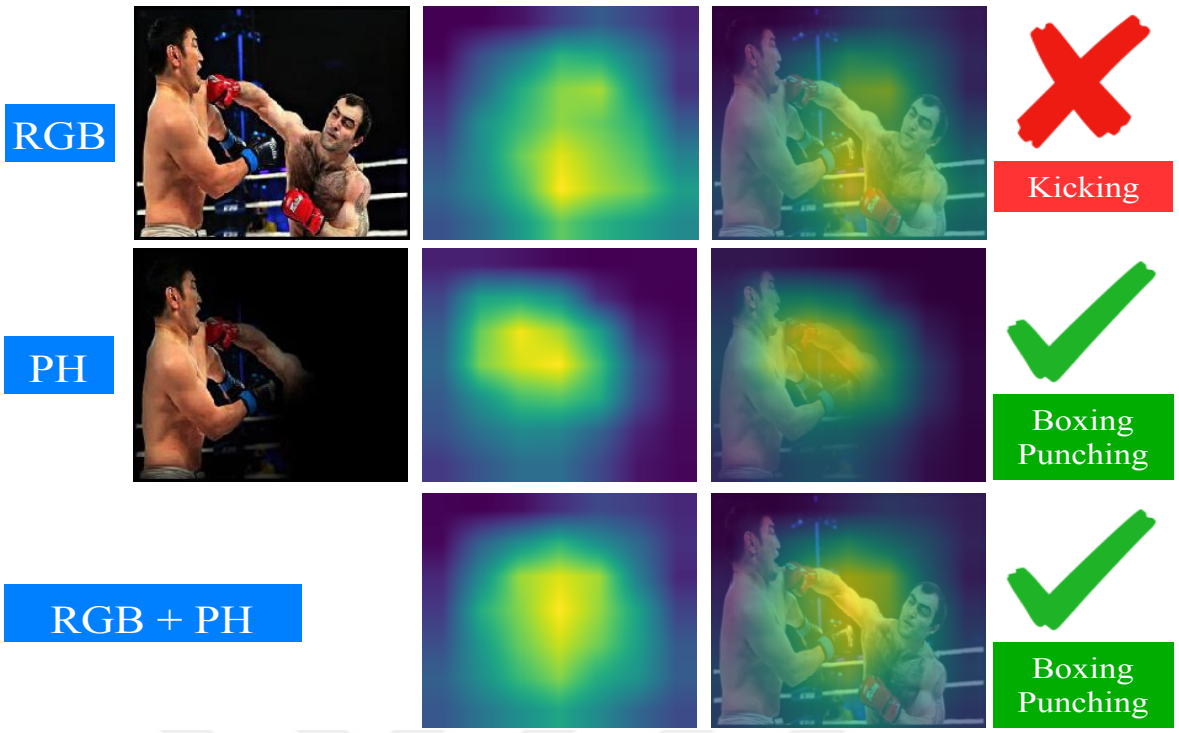


FIGURE 7.18: Visualization of network activations. `RGB, PH and RGB+PH` streams are presented for simplicity. Sample image from `Boxing-Punching` class.

Visualizations show that the model attends regions that are related to actual interactions. Moreover, `Pose Highlighted` model capable of finding essential spots. Another point is that the proposed fusion scheme `RGB + PH` enhances the salient regions and therefore performs better recognition performance.

7.2.2.7 Comparison to related work that applies to human-human interaction recognition.

Recognizing human-human interactions differs from recognizing action or human-object interactions. Action recognition requires to explore only one human's body composition. However, in most of the datasets, some of the actions also require an object to be performed. For instance, Stanford 40 Action [97] dataset defines actions such as *brushing teeth, cutting vegetables, holding an umbrella, riding a bike, etc*. Human-object interactions, on the other hand, always defined with one person interacting with a single object. Only highlighting a human body would lead to a misleading bias on the model. Regarding this, the proposed method is only deployed for the human-human interaction recognition problem. However a few works such as *Attentional Pooling* [104], *Human-mask Loss*[108], *Im2Flow* [127], and *Attention Branch Network (ABN)* [107] on action recognition can also be applied to human-human interactions. These approaches are deployed on *HIIv2* dataset and the results are shown in Table 7.13. Moreover, to create a baseline for future work, the contribution of

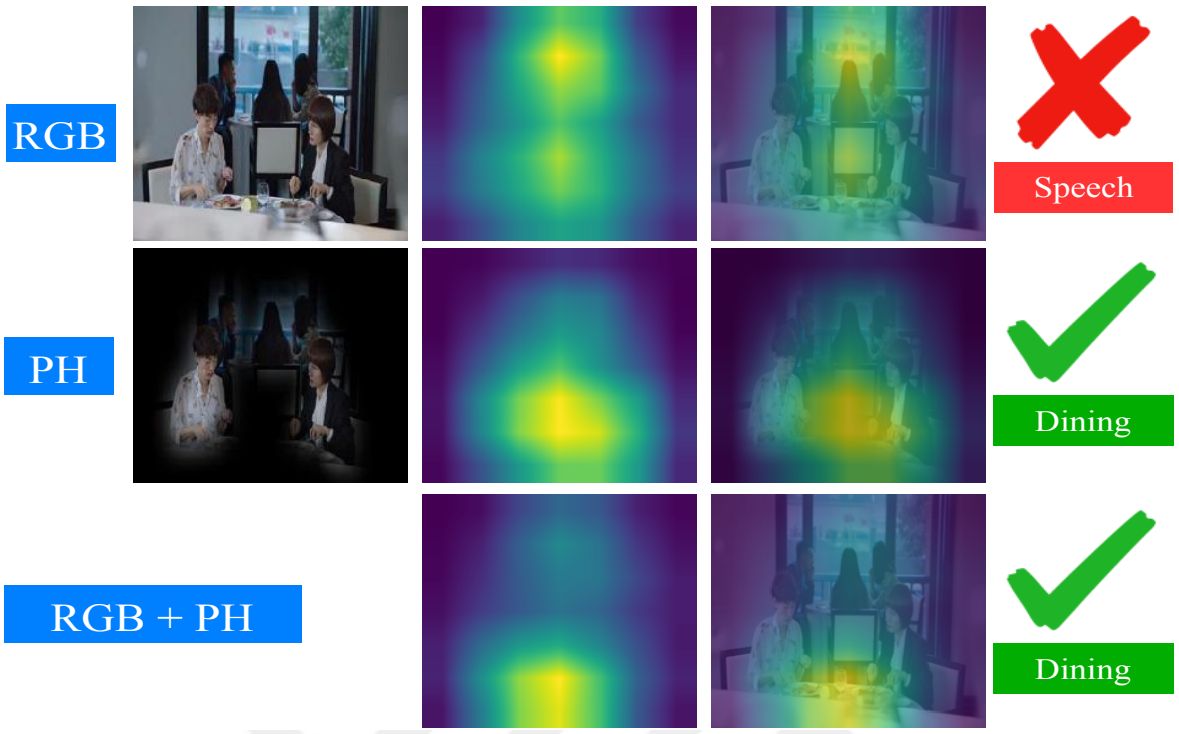


FIGURE 7.19: Plot of ResNet152 model Accuracy and Loss on both train and validation sets for the `RGB` stream. Sample image from `Dining` class.

FIGURE 7.20: Plot of ResNet152 model Accuracy and Loss on both train and validation sets for the `RGB` stream. Sample image from `HighFive` class.

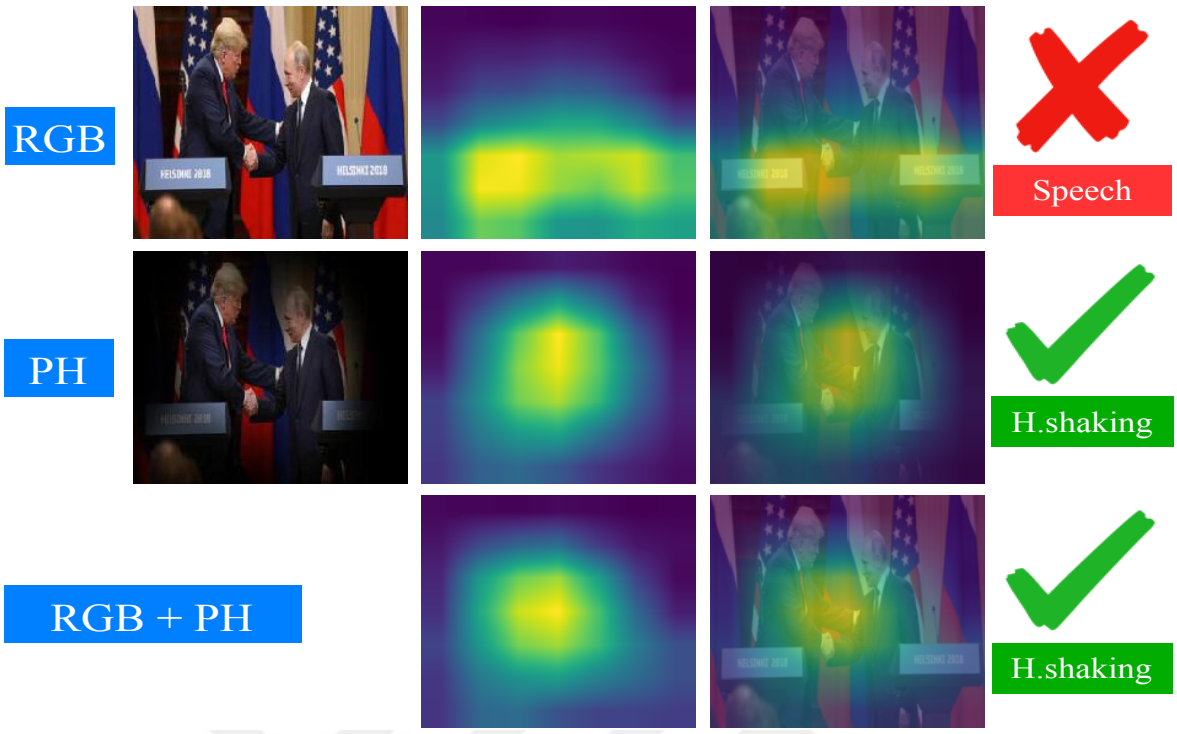


FIGURE 7.21: Plot of ResNet152 model Accuracy and Loss on both train and validation sets for the `RGB` stream. Sample image from `Handshaking` class.

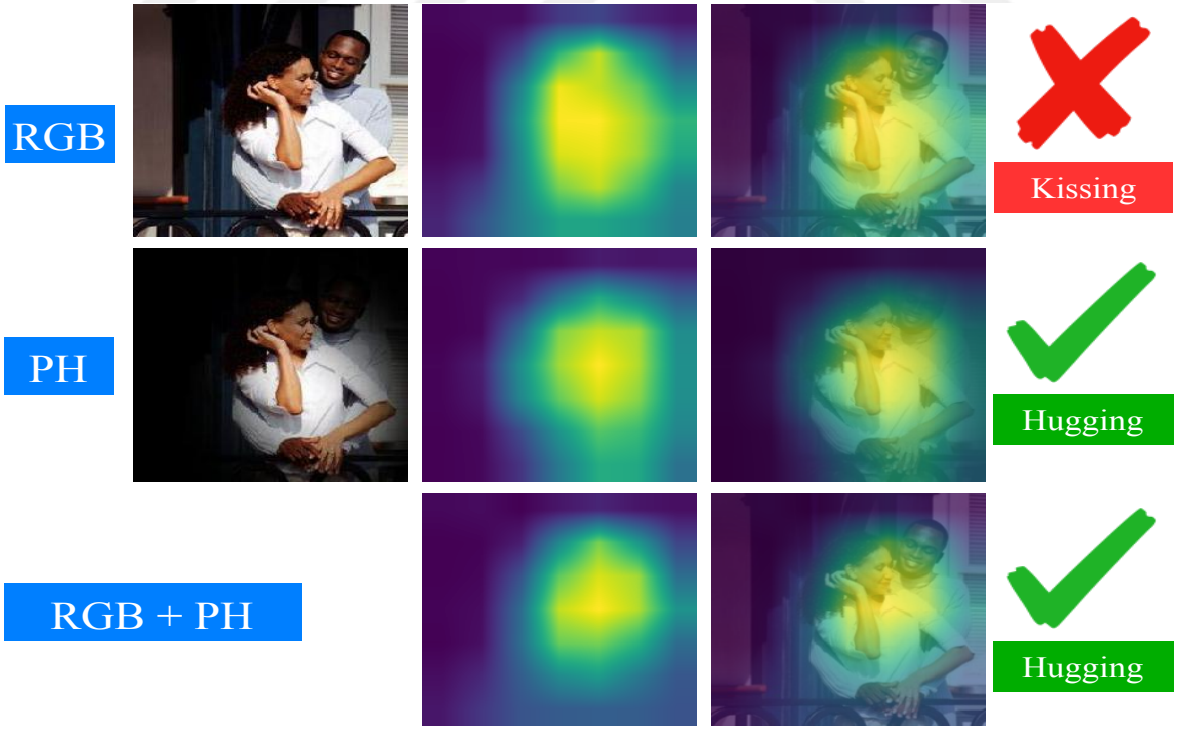


FIGURE 7.22: Visualization of network activations. `RGB, PH and RGB+PH` streams are presented for simplicity. Sample image from `Hugging` class.

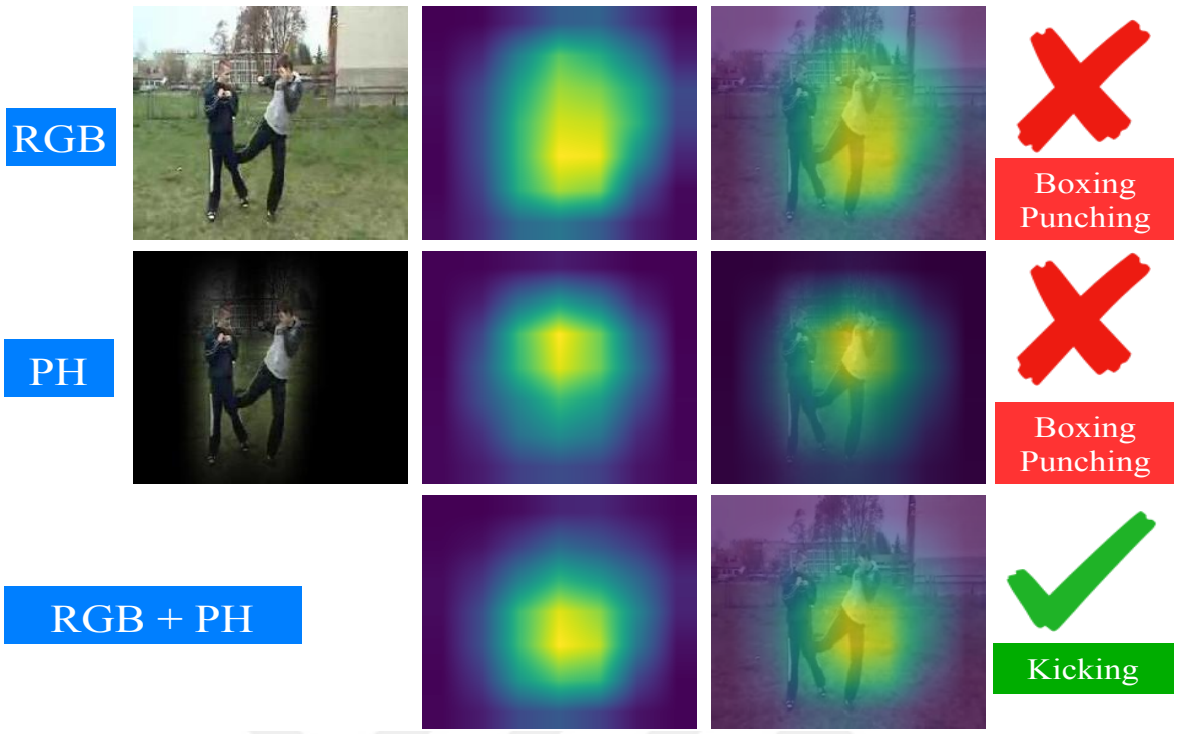


FIGURE 7.23: Visualization of network activations. `RGB, PH and RGB+PH` streams are presented for simplicity. Sample image from `Kicking` class.

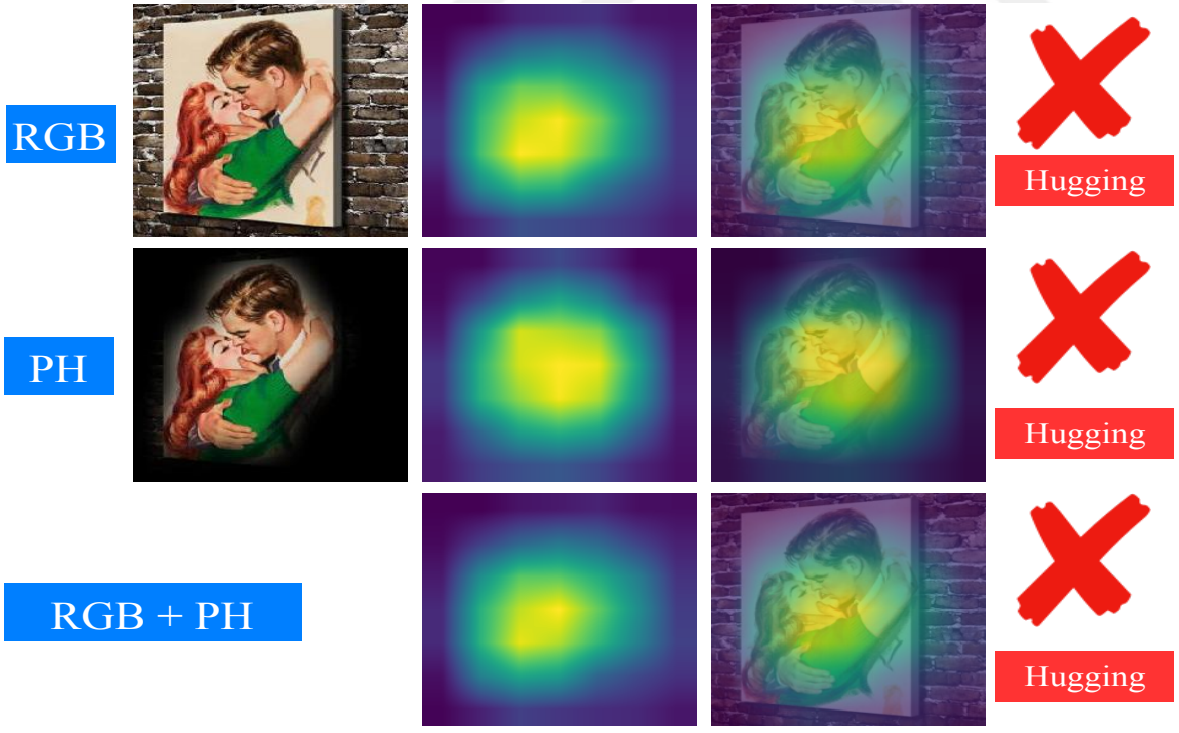


FIGURE 7.24: Visualization of network activations. `RGB, PH and RGB+PH` streams are presented for simplicity. Sample image from `Kissing` class.

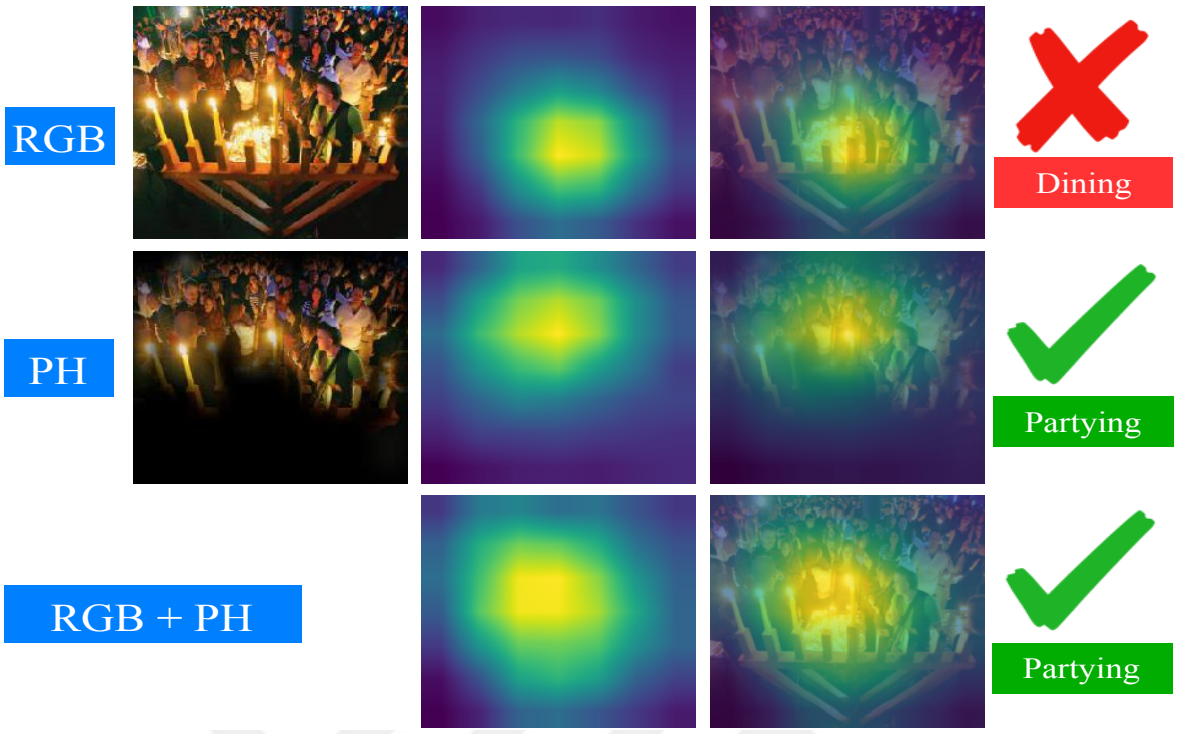


FIGURE 7.25: Visualization of network activations. `RGB, PH and RGB+PH` streams are presented for simplicity. Sample image from `Partying` class.

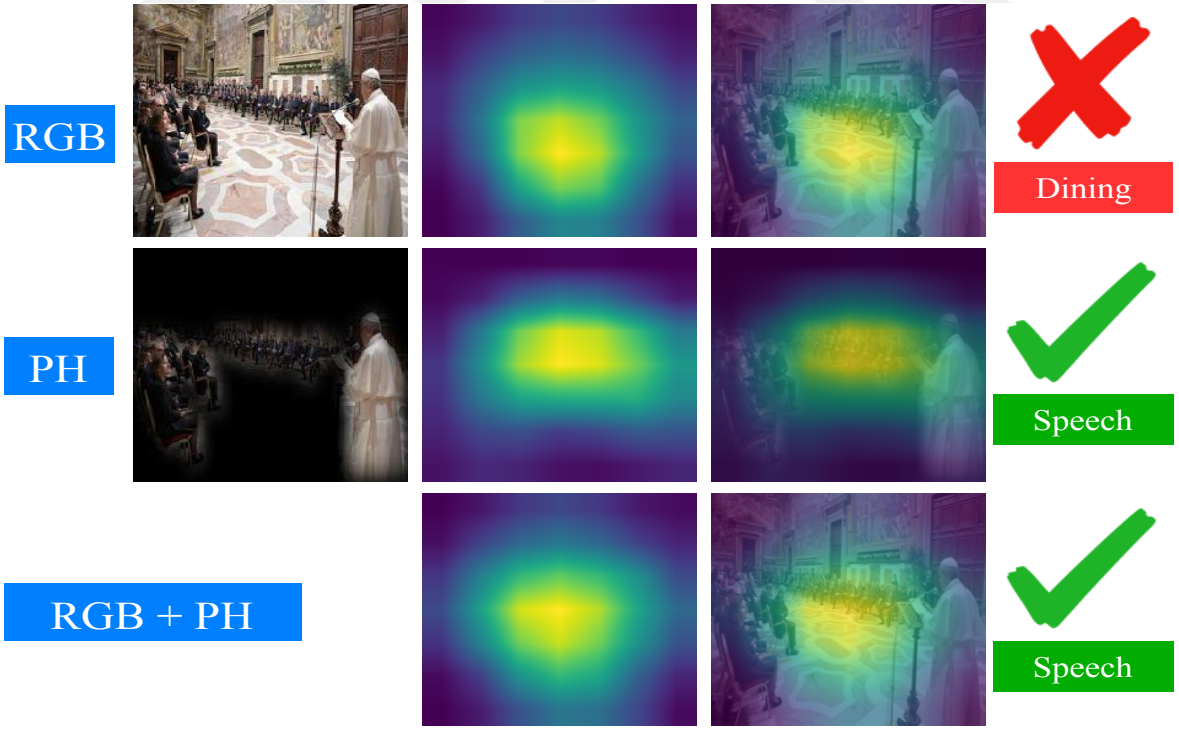


FIGURE 7.26: Visualization of network activations. `RGB, PH and RGB+PH` streams are presented for simplicity. Sample image from `Speech` class.

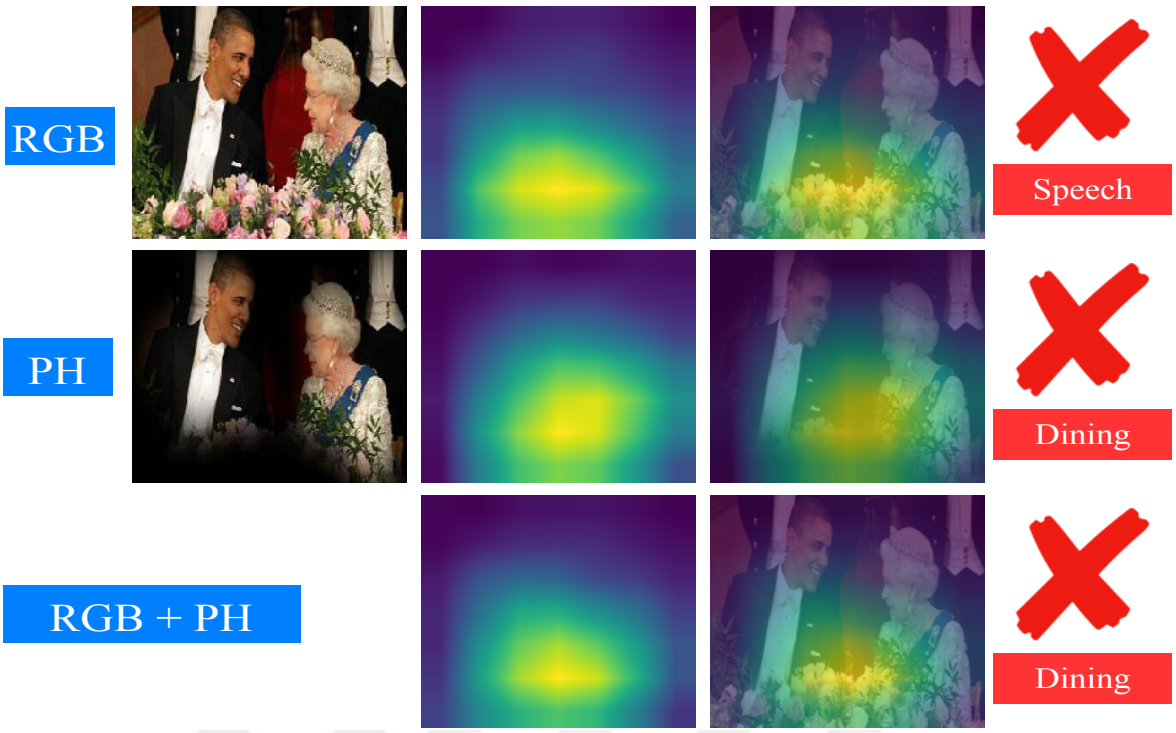


FIGURE 7.27: Visualization of network activations. `RGB, PH and RGB+PH` streams are presented for simplicity. Sample image from `Talking` class.

depths are evaluated by using the recent depth estimation method Monodepth2 [128] and the results are given in Table 7.13.

Attentional Pooling is shown to be a successful method for both action and human-object interaction. The basic idea is to change the classifier layer with a saliency-based attention module. With a small adaptation, it is easily trained on *HIIv2* dataset. However, it uses TensorFlow as the deep learning framework. The difference in two implementations yields different results on the same data. To make a fair comparison, along with the Attentional Pooling method, its baseline implementation trained on *HIIv2* is also presented in the table 7.13. For the *Pose Regularized Attentional Pooling* version, keypoints of the human pose are detected by AlphaPose [129].

Human-mask Loss is another method that also applies to human-interaction recognition. Its source code is not available online. Therefore it is implemented from scratch. [1]

Im2Flow [127] estimates optical flow images and use them together with the original input images to form a two-stream convolutional neural network architecture. The model is trained

[1] In order to make it directly comparable to the proposed work, same input image size ($224x224$) is used.

TABLE 7.13: Comparison of related methods applied to Human-Human Interaction.

Model	Accuracy	Δ Accuracy	mAP	Δ mAP
Baseline [104]	92.54	—	96.89	—
Attentional Pooling [104]	92.49	-0.05	96.89	0.09
Pose Regularized Attentional Pooling [104]	92.02	-0.52	96.63	-0.26
Baseline (This Work)	90.82	—	94.63	—
Human-mask Loss [108]	91.63	0.41	94.96	0.33
Attention Branch Network [107]	91.11	0.29	94.76	0.13
Flow [127]	62.91	-	65.33	-
Monodepth2 [128]	56.26	-	58.35	-
`RGB + Flow` [127]	91.44	0.62	95.15	0.52
`RGB + Monodepth2` [128]	91.73	0.91	95.73	1.10
`RGB + PM`	91.30	0.48	95.23	0.60
`RGB + PO`	92.11	1.29	95.94	1.31
`RGB + PH`	92.97	**2.15**	96.49	**1.86**
`RGB + PM + PO + PH`	91.54	0.72	94.31	-0.32

to detect human actions from still images. The same idea could be applied to interaction recognition, as well. First, optical flow hallucinations are extracted from *HIIv2* dataset and a two-stream network is similar to *RGB + Pose Highlighted* model.

Attention Branch Network [107] was used to compare the proposed model with attention-based models. This network makes a small modification on regular ResNet architecture, and it is directly applicable to a regular single-stream model. *Attentional Pooling* uses 101 layers ResNet (ResNet101) as the backbone CNN architecture. For the sake of consistency, Table 7.13 presents evaluations of the five approaches using ResNet101 on *HIIv2*. The results show that all models, except Attentional Pooling, have a positive impact over the baseline. It has been observed that attending to the salient regions alone are not useful for interactions. On the other hand Attentional Branch Network (ABN) [107] has a small positive impact on accuracy. This is an expected outcome since ABN learns where to attend during the training phase.

Surprisingly, optical flow estimation does an excellent job and leads to a boost on both accuracy and mAP. Human-mask Loss [108] method learns both the interactions and pose masks at the same time. This enables pose-aware learning of interactions. Among all, the proposed method demonstrates superior improvement. Comparisons presented in Table 7.13 show that concentrating on poses has a noticeable impact on interaction recognition performance.

TABLE 7.14: Evaluation of the approach with ResNet152 based on different human segmentation algorithms.

Model	MaskRCNN[1]	DeepMask[130]	SharpMask [60]	DeepLab [59]
PM	**77.01**	39.75	43.44	74.43
PO	87.72	82.46	82.89	**88.31**
PH	90.34	88.13	87.74	**90.35**
RGB + PM	**92.45**	91.34	91.44	91.40
RGB + PO	92.02	90.90	91.54	**93.01**
RGB + PH	**92.97**	91.38	91.68	92.49
RGB + PM + PO	**91.87**	90.54	90.35	91.78
RGB + PM + PH	**93.26**	91.55	91.11	92.30
RGB + PO + PH	92.69.	90.63	91.25	**92.73**
RGB + PM + PO + PH	**92.50**	89.68	89.78	89.40

Another observation can be made by Table 7.13 is that both the proposed and *Human-mask Loss* [108] methods yields to similar results. This is an expected outcome since the basic idea in both approaches is the same: paying more attention to human bodies.

7.2.2.8 Impact of different segmentation algorithms

The accuracy of the proposed method increases with the success of the underlying segmentation algorithm. In order to evaluate this, four recent segmentation algorithms are utilized, namely DeepMask [130], SharpMask [60], MaskRCNN [1] and DeepLab [1] [59].

Table 7.14 shows that the success of the segmentation algorithm has an essential influence on the final results. Here, it was observed that the best results are obtained by MaskRCNN [1] overall. It should be noted that, even though MaskRCNN [1] is a state-of-the-art instance segmentation algorithm, there are cases of failure. With a better segmentation algorithm, the accuracy of the proposed method may increase further.

7.2.2.9 Impact of varying hyper-parameters on the results

Hyper-parameters of the models are optimized by conducting grid-search on the HIIv2 dataset for each network backbone individually. The grid search suffers from the curse of dimensionality. Moreover, there are too many model-network backbone-dataset tuples to be optimized. To overcome this issue, each network backbone is optimized on the HIIv2 dataset for one single-stream, and one of the two-stream models and the rest of the models share the

[1]To generate the Pose Highlighted representation for DeepLab, human body sizes are used as the same as MaskRCNN.

TABLE 7.15: Evaluation of hyper-parameter optimization via grid search on `Pose Highlighted` representation.

Learning Rate	Step Size	Gamma	Optimizer	SGD Momentum	Batch Size	Sampling	Accuracy	mAP
0.0005	10	0.25	SGD	0.9	8	Weighted	90.92	93.79
0.0005	10	0.25	SGD	0.9	8	*Random*	90.45	93.67
0.001	10	0.25	SGD	0.9	8	Weighted	90.11	93.40
0.0001	10	0.25	SGD	0.9	8	Weighted	89.02	92.84
0.00075	10	0.25	SGD	0.9	8	Weighted	90.35	93.56
0.0005	10	0.25	*Adam*	—	8	Weighted	82.75	83.90
0.0005	*7*	0.25	SGD	0.9	8	Weighted	90.35	93.57
0.0005	10	*0.5*	SGD	0.9	8	Weighted	90.87	93.84
0.0005	10	0.25	SGD	*0.5*	8	Weighted	89.26	92.87
0.0005	10	0.25	SGD	0.9	*16*	Weighted	90.68	93.51

same parameters. Besides, only the most critical parameters are optimized to reduce the size of the parameter space.

The results with varying hyper-parameters are presented in Table 7.15 and Table 7.16. Respectively, Table 7.15 shows the results of parameter search process on single `Pose Highlighted` stream and Table 7.16 shows the results for `RGB + Pose Highlighted`. The first row of each table shows the best performing parameter set on the validation set. In the rest, all the parameters are fixed except one. In each row, the changing parameter is shown with italic to read the table easier. The *Step Size* in the tables refers to the period of learning rate decay, and *gamma* refers to the multiplicative factor of learning rate decay for the learning rate update policy. Although the number of instances in each class of *HIIv2* is well balanced (please refer to Table 6.1), the distribution is not equal. Therefore, two sampling strategies are considered: *random sampling* and *weighted random sampling*.

Based on the results, we observe that *learning rate* and *SGD momentum* are the most effective hyperparameters on the performance. Batch size is also effective on the performance because of *batch normalization*. The results indicate that using weighted sampling yields better results and that a parameter optimization step is crucial which should be preferred over empirical parameter selection.

7.3 Comparison of Facial Descriptors and Multi-stream Pose CNNs

In a perfect scenario, to accurately detect an interaction, we would need to see the whole sequence, everyone involved, the whole context, and even hear what they talk to. However,

TABLE 7.16: Evaluation of hyper-parameter optimization via grid search on `RGB + Pose Highlighted` representation.

Learning Rate	Step Size	Gamma	Optimizer	SGD Momentum	Batch Size	Sampling	Accuracy	mAP
0.00025	10	0.25	SGD	0.9	10	Weighted	92.82	95.87
0.00025	10	0.25	SGD	0.9	10	*Random*	92.63	96.01
0.0001	10	0.25	SGD	0.9	10	Weighted	92.30	95.96
0.0005	10	0.25	SGD	0.9	10	Weighted	92.44	95.80
0.00025	10	0.25	*Adam*	—	10	Weighted	88.36	89.36
0.00025	*7*	0.25	SGD	0.9	10	Weighted	92.54	96.09
0.00025	10	*0.5*	SGD	0.9	10	Weighted	92.54	96.09
0.00025	10	0.25	SGD	*0.5*	10	Weighted	92.25	95.99
0.00025	10	0.25	SGD	0.9	*8*	Weighted	92.78	95.70

TABLE 7.17: Comparison of the works based on faces and poses.

	Model	mAP	$\Delta Boost$
FaceDesc	HybridCNN + SVM	77.86	—
	(HybridCNN + FaceDesc) + SVM	78.33	0.47
Pose CNNs	`RGB`	72.66	—
	`RGB + PM`	82.01	9.35
	`RGB + PO`	81.13	8.47
	`RGB + PH`	79.52	6.86
	`RGB + PM + PO + PH`	**83.73**	**11.07**

in most cases, an image with the context would be sufficient to describe the interaction. We propose two different approaches to recognize human interactions in still images. One explores the role of human faces, and the other explores the role of human poses. Although faces would tell some about the interaction, they cannot capture all the information (see Figure 1.1). Therefore it would not be fair to compare these two methods. It is obvious that poses would perform better than faces.

Having this in mind, comparison results are given on Table 7.17. The work based on facial descriptors uses HybridCNN [116], which basically is Alexnet [3] trained on a union of Places [116] and Imagenet [45] datasets. HybridCNN is used to extract features and SVM is used for classification. Moreover, hand-crafted facial features are extracted to describe interactions, such as positions of human faces in the image. In order to be directly comparable, we show classification results of pose representations using the same Alexnet [3] architecture, trained on Imagenet and fine-tuned on HII dataset. Accuracy and mAP scores are shown in Table 7.17. The results show that each of the pose streams makes a greater boost

in performance and they contribute better than faces when the multi-stream pose CNNs are used.

8. CONCLUSION AND DISCUSSIONS

Two methods to recognize human interactions in images are explored in this thesis. In the first part, we examine the role of human faces in the images. We look into a rarely studied area of computer vision, namely human-human interaction recognition in still images. We investigate whether we can infer the correct label of an interaction image by looking at the facial regions, their relative positions, and spatial layout. In order to capture such information, we propose several descriptors based on facial regions. The proposed facial descriptors encode spatial information such as their position or distance to each other. Our experimental results show that facial descriptors provide meaningful information; however, using them in isolation yields less effective results. When combined with global scene features, especially deep features, proposed facial descriptors are shown to have improved recognition performance.

In this context, we introduce a new image dataset, Human Interaction Images (HII), which can be used for human interaction recognition, and also for evaluating face detectors' performance on further tasks. There exist ten interaction classes in the dataset: *boxing-punching, dining, handshaking, high five, hugging, kicking, kissing, partying, speech*, and *talking*. Each class contains at least 150 images, forming a total of 1971 images. The faces in the dataset are annotated with both locations and orientations and published online.

In the second part of the thesis, the role of poses in classifying human-human interactions in still images is examined. To this end, multi-stream pose CNNs are introduced, where each stream operates over different human pose representations. Three different pose input variations are proposed, namely, pose masks, pose only images, and pose highlighted images. To evaluate the proposed approach, we've collected *HIIv2* dataset that is an extension of *HII* dataset. *HIIv2* contains the same number of classes, 10, with more than 10000 images in total. It is shown that human poses, without the full context, are powerful enough to express an interaction.

A multi-stream Convolutional Neural Network model is proposed to force the CNN model to attend more on the human poses. Our pose-based inputs act as an intuitive spatial attention model for the original CNN architecture. Experimental results show that each of the proposed

representations provides valuable information to classify an interaction. A deep neural network has an extreme capacity for classifying interactions. Nevertheless, it is shown that with correct guidance, it has more to improve.

Extensive experiments on both existing (*HII*) and newly extended (*HIIv2*) are conducted for evaluations. Well known CNN architectures are examined and shown that recognition performance rises with the depth of the model. A comparison to the previous approach is given and revealed that poses are more important than human faces in recognizing interactions.

Finally, existing methods to classify human actions and human-object interactions are applied to human-human interaction recognition and evaluated on *HIIv2* dataset. It is observed that attending the salient regions does not help much. On the other hand, attending poses is shown to provide a noticeable contribution to classifying interactions.

To sum up, we showed that our methods are capable of recognizing human interactions from still images. A remarkable result is that using only human faces, we may have a clue of the interaction in an image. Another result to emphasize is that just human body masks alone are capable of describing the interaction. This is surprising because the input is a binary image, where the masks are not even perfect. Furthermore, human bodies alone are shown to be the essential component of an interaction, even without motion, they almost entirely describe an interaction. Finally, poses combined with original input images tend to increase the performance of recognition. This helps the network to give more attention to human bodies, which increases accuracy.

Various methods are evaluated during the study of this thesis. However, there are some aspects which may have further contributions. One aspect would be further evaluations of faces, such as the emotions of people in the image scene. Emotion recognition would be helpful to discriminate interactions. For example, a person being punched would be suffering pain, and people in a party would be happy. Emotion recognition is an active topic of computer vision and novel methods [131, 132] have advancements, and it may have promising results on interaction recognition.

Another aspect would be using depths more effectively to make use of the most relevant information in the scene. Using depths enables the opportunity of ignoring irrelevant people

or objects in the scene. Moreover, it becomes feasible to separate interactions in a scene. Thus, recognizing multiple interactions in a scene would be possible. A baseline setup is created for depths, and evaluations are given in Table 7.13. More detailed work on depths has promising contributions to the recognition of interactions.

In our method, we fused different instances of the same architecture for the proposed models. Candidate future work methods, such as depth estimation and emotion recognition, would have very different architectures which makes it harder to use a simple fusion scheme. A Graph Neural Network [133] stands as a compelling choice and is considered to be useful for future work.

The focus of this thesis is recognizing the most obvious and notable interaction in an image. As a future work, the below problems are open to exploring:

- Localizing the interaction. We have answered the question of which interaction is being performed. However, we ignore where exactly it is being performed. Finding the location of interaction would be an interesting subject to solve.
- Recognizing multiple interactions in a scene. This could be considered in two different ways. First, some interactions may naturally involve another one. For example, people kissing may also be hugging each other. The second one is, depending on the field of view, there might be multiple interactions occurring in the scene. Thus, a recognition system should consider each interaction individually.
- Finding the actors of an interaction. Some interactions require active and passive actors to perform. For example, kicking requires an active person performing the action and a passive person receiving the kick. Thus, detecting the actors is a problem to solve. Moreover, in most cases, there are redundant people in a scene who do not relate to the interaction. Finding these people may increase recognition performance.

REFERENCES

[1] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988. **2017**. doi:10.1109/ICCV.2017.322.

[2] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2879–2886. **2012**. ISSN 1063-6919. doi:10.1109/CVPR.2012.6248014.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pages 1097–1105. Curran Associates Inc., USA, **2012**.

[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. **2016**. doi:10.1109/CVPR.2016.90.

[5] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. pages 2261–2269, **2020**.

[6] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *International Conference on Neural Information Processing Systems - Volume 1*, NIPS'14, pages 568–576. MIT Press, Cambridge, MA, USA, **2014**.

[7] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1933–1941. **2016**. doi:10.1109/CVPR.2016.213.

[8] Yu Kong and Yun Fu. Human action recognition and prediction: A survey. *CoRR*, abs/1806.11230, **2018**.

[9] Alexandros Stergiou and Ronald Poppe. Understanding human-human interactions: a survey. *CoRR*, abs/1808.00022, **2018**.

[10] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221–231, **2013**. doi:10.1109/TPAMI.2012.59.

[11] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):677–691, **2017**. doi:10.1109/TPAMI.2016.2599174.

[12] Moez Baccouche, Franck Mamalet, Christian Wolf, Christophe Garcia, and Atilla Baskurt. Sequential deep learning for human action recognition. In Albert Ali Salah and Bruno Lepri, editors, *Human Behavior Understanding*, pages 29–39. Springer Berlin Heidelberg, Berlin, Heidelberg, **2011**. ISBN 978-3-642-25446-8. doi:10.1007/978-3-642-25446-8_4.

[13] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1725–1732. **2014**. doi:10.1109/CVPR.2014.223.

[14] Limin Wang Luc Van Gool Yifan Wang, Jie Song and Otmar Hilliges. Two-stream sr-cnns for action recognition in videos. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 108.1–108.12. BMVA Press, **2016**. ISBN 1-901725-59-6. doi:10.5244/C.30.108.

[15] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733. **2017**. doi:10.1109/CVPR.2017.502.

[16] Waleed Abdulla. Mask r-cnn for object detection and instance segmentation on keras and tensorflow, **2017**.

[17] Theodore M. Newcomb, Ralph H. Turner, and Philip E. Converse. *Social psychology : the study of human interaction*. Tavistock Publications (in collaboration with) Routledge and Kegan Paul London, 2nd ed. (rev.) edition, **1966**. ISBN 0710018908.

[18] G. Varol, I. Laptev, and C. Schmid. Long-term temporal convolutions for action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1510–1517, **2018**. doi:10.1109/TPAMI.2017.2712608.

[19] R. Girdhar, D. Ramanan, A. Gupta, J. Sivic, and B. Russell. Actionvlad: Learning spatio-temporal aggregation for action classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3165–3174. **2017**. doi:10.1109/CVPR.2017.337.

[20] M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori. A hierarchical deep temporal model for group activity recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1971–1980. **2016**. doi:10.1109/CVPR.2016.217.

[21] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *AAAI Conference on Artificial Intelligence*, AAAI'17, pages 4263–4270. AAAI Press, **2017**.

[22] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 407–423. Springer International Publishing, Cham, **2018**. ISBN 978-3-030-01240-3. doi:10.1007/978-3-030-01240-3_25.

[23] Y. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng. Learning to detect human-object interactions. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 381–389. **2018**. doi:10.1109/WACV.2018.00048.

[24] Arun Mallya and Svetlana Lazebnik. Learning models for actions and person-object interactions with transfer to question answering. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 414–428. Springer International Publishing, Cham, **2016**.

[25] B. Yao and L. Fei-Fei. Grouplet: A structured image representation for recognizing human and object interactions. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 9–16. **2010**. doi:10.1109/CVPR.2010.5540234.

[26] Yixiao Yun, I.Y.-H. Gu, and H. Aghajan. Riemannian manifold-based support vector machine for human activity classification in images. In *IEEE International Conference on Image Processing*, pages 3466–3469. **2013**. doi:10.1109/ICIP.2013.6738715.

[27] A. Fathi, J. K. Hodgins, and J. M. Rehg. Social interactions: A first-person perspective. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1226–1233. **2012**. ISSN 1063-6919. doi:10.1109/CVPR.2012.6247805.

[28] M. S. Ryoo and J. K. Aggarwal. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In *International Conference on Computer Vision*, pages 1593–1600. **2009**. doi:10.1109/ICCV.2009.5459361.

[29] M. S. Ryoo and J. K. Aggarwal. UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA), **2010**.

[30] Alonso Patron, Marcin Marszalek, Andrew Zisserman, and Ian Reid. High five: Recognising human interactions in tv shows. In *British Machine Vision Conference*, pages 50.1–50.11. BMVA Press, **2010**. ISBN 1-901725-40-5. doi:10.5244/C.24.50.

[31] B. Ni, Gang Wang, and P. Moulin. Rgbd-hudaact: A color-depth video database for human daily activity recognition. In *IEEE International Conference*

on Computer Vision Workshops (ICCV Workshops), pages 1147–1153. **2011**. doi:10.1109/ICCVW.2011.6130379.

[32] M. S. Ryoo and L. Matthies. First-person activity recognition: What are they doing to me? In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2730–2737. **2013**. doi:10.1109/CVPR.2013.352.

[33] F. Siyahjani, S. Motiian, H. Bharthavarapu, S. Sharlemin, and G. Doretto. Online geometric human interaction segmentation and recognition. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. **2014**. doi:10.1109/ICME.2014.6890330.

[34] R. Li, P. Porfilio, and T. Zickler. Finding group interactions in social clutter. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2722–2729. **2013**. doi:10.1109/CVPR.2013.351.

[35] S. Motiian, F. Siyahjani, R. Almohsen, and G. Doretto. Online human interaction detection and recognition with multiple cameras. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(3):649–663, **2017**. doi:10.1109/TCSVT.2016.2606998.

[36] Wenjuan Gong, Jordi Gonzàlez, João Manuel R. S. Tavares, and F. Xavier Roca. A new image dataset on human interactions. In Francisco J. Perales, Robert B. Fisher, and Thomas B. Moeslund, editors, *Articulated Motion and Deformable Objects*, pages 204–209. Springer Berlin Heidelberg, Berlin, Heidelberg, **2012**. ISBN 978-3-642-31567-1. doi:10.1007/978-3-642-31567-1_20.

[37] A. C. Gallagher and T. Chen. Understanding images of groups of people. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 256–263. **2009**. doi:10.1109/CVPR.2009.5206828.

[38] Y. Yang, S. Baker, A. Kannan, and D. Ramanan. Recognizing proxemics in personal photos. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3522–3529. **2012**. doi:10.1109/CVPR.2012.6248095.

[39] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, **1995**. ISSN 1573-0565. doi:10.1023/A:1022627411411.

[40] Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, and Dan Steinberg. Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1):1–37, **2008**. ISSN 0219-3116. doi:10.1007/s10115-007-0114-2.

[41] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893 vol. 1. **2005**. doi:10.1109/CVPR.2005.177.

[42] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, **2001**. ISSN 1573-1405. doi:10.1023/A:1011139631724.

[43] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, **2004**. ISSN 0920-5691. doi:10.1023/B:VISI.0000029664.99615.94.

[44] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 2169–2178. **2006**. ISSN 1063-6919. doi:10.1109/CVPR.2006.68.

[45] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, **2015**. ISSN 1573-1405. doi:10.1007/s11263-015-0816-y.

[46] Yann Lecun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 2278–2324. **1998**.

[47] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, **2014**.

[48] C. Szegedy, , , P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9. **2015**. ISSN 1063-6919. doi:10.1109/CVPR.2015.7298594.

[49] Guanyu Li, Xiaoling Xu, Minghui Zhang, and Qiegen Liu. Densely connected network for impulse noise removal. In *Pattern Analysis and Applications*. **2017**.

[50] E. Park, X. Han, T. L. Berg, and A. C. Berg. Combining multiple sources of knowledge in deep cnns for action recognition. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–8. **2016**. doi:10.1109/WACV.2016.7477589.

[51] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, **2017**. ISSN 0162-8828. doi:10.1109/TPAMI.2016.2577031.

[52] J.K. Aggarwal and M.S. Ryoo. Human activity analysis: A review. *ACM Comput. Surv.*, 43(3):16:1–16:43, **2011**. ISSN 0360-0300. doi:10.1145/1922649.1922653.

[53] I. Chakraborty, H. Cheng, and O. Javed. 3d visual proxemics: Recognizing human interactions in 3d from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3406–3413. **2013**. doi:10.1109/CVPR.2013.437.

[54] L. Xia and J. K. Aggarwal. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2834–2841. **2013**. ISSN 1063-6919. doi:10.1109/CVPR.2013.365.

[55] W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3d points. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 9–14. **2010**. ISSN 2160-7508. doi:10.1109/CVPRW.2010.5543273.

[56] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras. Two-person interaction detection using body-pose features and multiple instance learning. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–35. **2012**. ISSN 2160-7508. doi:10.1109/CVPRW.2012.6239234.

[57] Y. Kong and Y. Fu. Bilinear heterogeneous information machine for rgb-d action recognition. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1054–1062. **2015**. ISSN 1063-6919. doi:10.1109/CVPR.2015.7298708.

[58] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *arXiv preprint arXiv:2001.05566*, **2020**.

[59] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 833–851. Springer International Publishing, Cham, **2018**.

[60] Pedro O. Pinheiro, Tsung-Yi Lin, Ronan Collobert, and Piotr Dollár. Learning to refine object segments. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 75–91. Springer International Publishing, Cham, **2016**.

[61] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea. Machine recognition of human activities: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(11):1473–1488, **2008**. ISSN 1051-8215. doi:10.1109/TCSVT.2008.2005594.

[62] P. V. K. Borges, N. Conci, and A. Cavallaro. Video-based human behavior understanding: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(11):1993–2008, **2013**. ISSN 1051-8215. doi:10.1109/TCSVT.2013.2270402.

[63] O. Yurur, C. H. Liu, and W. Moreno. A survey of context-aware middleware designs for human activity recognition. *IEEE Communications Magazine*, 52(6):24–31, **2014**. ISSN 0163-6804. doi:10.1109/MCOM.2014.6829941.

[64] Ronald Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6):976 – 990, **2010**. ISSN 0262-8856. doi:https://doi.org/10.1016/j.imavis.2009.11.014.

[65] Daniel Weinland, Remi Ronfard, and Edmond Boyer. A survey of vision-based methods for action representation, segmentation and recognition. *Computer Vision and Image Understanding*, 115(2):224 – 241, **2011**. ISSN 1077-3142. doi:https://doi.org/10.1016/j.cviu.2010.10.002.

[66] Hong-Bo Zhang, Yi-Xiang Zhang, Bineng Zhong, Qing Lei, Lijie Yang, Ji-Xiang Du, and Duan-Sheng Chen. A comprehensive survey of vision-based human action recognition methods. *Sensors*, 19(5), **2019**. ISSN 1424-8220. doi:10.3390/s19051005.

[67] Yu Kong and Yun Fu. *Action Recognition and Human Interaction*. Springer, Cham, **2016**. doi:https://doi.org/10.1007/978-3-319-27004-3.

[68] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):257–267, **2001**. ISSN 0162-8828. doi:10.1109/34.910878.

[69] Manuel J. MarÃn-JimÃ©nez, Enrique Yeguas, and NicolÃ¡s PÃ©rez de la Blanca. Exploring stip-based models for recognizing human interactions in tv videos. *Pattern Recognition Letters*, 34(15):1819 – 1828, **2013**. ISSN 0167-8655. doi:https://doi.org/10.1016/j.patrec.2012.10.018. Smart Approaches for Human Action Recognition.

[70] B. Zhang, F. G. B. De Natale, and N. Conci. Recognition of social interactions based on feature selection from visual codebooks. In *2013 IEEE International Conference on Image Processing*, pages 3557–3561. **2013**. ISSN 1522-4880. doi:10.1109/ICIP.2013.6738734.

[71] Ju Sun, Xiao Wu, Shuicheng Yan, L. Cheong, T. Chua, and Jintao Li. Hierarchical spatio-temporal context modeling for action recognition. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2004–2011. **2009**. ISSN 1063-6919. doi:10.1109/CVPR.2009.5206721.

[72] Laptev and Lindeberg. Space-time interest points. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 432–439 vol.1. **2003**. doi:10.1109/ICCV.2003.1238378.

[73] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision*, 103(1):60–79, **2013**. ISSN 1573-1405. doi:10.1007/s11263-012-0594-8.

[74] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. **2008**. ISSN 1063-6919. doi:10.1109/CVPR.2008.4587756.

[75] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 3, pages 32–36 Vol.3. **2004**. ISSN 1051-4651. doi:10.1109/ICPR.2004.1334462.

[76] Xiaofei Ji, Changhui Wang, and Zhaojie Ju. A new framework of human interaction recognition based on multiple stage probability fusion. *Applied Sciences*, 7(6), **2017**. ISSN 2076-3417. doi:10.3390/app7060567.

[77] L. Morency, A. Quattoni, and T. Darrell. Latent-dynamic discriminative models for continuous gesture recognition. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. **2007**. ISSN 1063-6919. doi:10.1109/CVPR.2007.383299.

[78] X. Wu, D. Xu, L. Duan, and J. Luo. Action recognition using context and appearance distribution features. In *CVPR 2011*, pages 489–496. **2011**. ISSN 1063-6919. doi:10.1109/CVPR.2011.5995624.

[79] C. Yuan, X. Li, W. Hu, H. Ling, and S. J. Maybank. Modeling geometric-temporal context with directional pyramid co-occurrence for action recognition. *IEEE Transactions on Image Processing*, 23(2):658–672, **2014**. ISSN 1057-7149. doi:10.1109/TIP.2013.2291319.

[80] S. Wong, T. Kim, and R. Cipolla. Learning motion categories using both semantic and structural information. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–6. **2007**. ISSN 1063-6919. doi:10.1109/CVPR.2007.383332.

[81] Juan Carlos Niebles, Hongcheng Wang, and Li Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision*, 79(3):299–318, **2008**. ISSN 1573-1405. doi:10.1007/s11263-007-0122-4.

[82] Kui Jia and Dit-Yan Yeung. Human action recognition using local spatio-temporal discriminant embedding. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. **2008**. ISSN 1063-6919. doi:10.1109/CVPR.2008.4587732.

[83] A. Klaeser, M. Marszalek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *Proceedings of the British Machine Vision Conference*, pages 99.1–99.10. BMVA Press, **2008**. ISBN 1-901725-36-7. Doi:10.5244/C.22.99.

[84] S. Mohammadi, H. Kiani, A. Perina, and V. Murino. Violence detection in crowded scenes using substantial derivative. In *2015 12th IEEE International*

Conference on Advanced Video and Signal Based Surveillance (AVSS), pages 1–6. **2015**. doi:10.1109/AVSS.2015.7301787.

[85] G. Luo, S. Yang, G. Tian, C. Yuan, W. Hu, and S. J. Maybank. Learning human actions by combining global dynamics and local appearance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(12):2466–2482, **2014**. ISSN 0162-8828. doi:10.1109/TPAMI.2014.2329301.

[86] C. Yuan, W. Hu, G. Tian, S. Yang, and H. Wang. Multi-task sparse learning with beta process prior for action recognition. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 423–429. **2013**. ISSN 1063-6919. doi:10.1109/CVPR.2013.61.

[87] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497. **2015**. doi:10.1109/ICCV.2015.510.

[88] L. Wang, Y. Qiao, and X. Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4305–4314. **2015**. doi:10.1109/CVPR.2015.7299059.

[89] R. Girdhar, D. Ramanan, A. Gupta, J. Sivic, and B. Russell. Actionvlad: Learning spatio-temporal aggregation for action classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3165–3174. **2017**. doi:10.1109/CVPR.2017.337.

[90] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, **1997**. doi:10.1162/neco.1997.9.8.1735.

[91] C. J. Taylor. Reconstruction of articulated objects from point correspondences in a single uncalibrated image. In *IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No.PR00662)*, volume 1, pages 677–684 vol.1. **2000**. doi:10.1109/CVPR.2000.855885.

[92] Yang Wang, Hao Jiang, M. S. Drew, Ze-Nian Li, and G. Mori. Unsupervised discovery of action classes. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1654–1661. **2006**. doi:10.1109/CVPR.2006.321.

[93] L. Li and Li Fei-Fei. What, where and who? classifying events by scene and object recognition. In *International Conference on Computer Vision*, pages 1–8. **2007**. doi:10.1109/ICCV.2007.4408872.

[94] N. Ikizler, R. G. Cinbis, S. Pehlivan, and P. Duygulu. Recognizing actions from still images. In *International Conference on Pattern Recognition*, pages 1–4. **2008**. doi:10.1109/ICPR.2008.4761663.

[95] C. Thurau and V. Hlavac. Pose primitive based human action recognition in videos or still images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. **2008**. doi:10.1109/CVPR.2008.4587721.

[96] Vincent Delaitre, Ivan Laptev, and Josef Sivic. Recognizing human actions in still images: a study of bag-of-features and part-based representations. In *British Machine Vision Conference*, pages 97.1–97.11. BMVA Press, **2010**. ISBN 1-901725-40-5. doi:10.5244/C.24.97.

[97] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *International Conference on Computer Vision*, pages 1331–1338. **2011**. doi:10.1109/ICCV.2011.6126386.

[98] Y. Zheng, Y. Zhang, X. Li, and B. Liu. Action recognition in still images using a combination of human pose and context information. In *IEEE International Conference on Image Processing*, pages 785–788. **2012**. doi:10.1109/ICIP.2012.6466977.

[99] K. Raja, I. Laptev, P. Pérez, and L. Oisel. Joint pose estimation and action recognition in image graphs. In *IEEE International Conference on Image Processing*, pages 25–28. **2011**. doi:10.1109/ICIP.2011.6116197.

[100] Fadime Sener, Cagdas Bas, and Nazli Ikizler-Cinbis. On recognizing actions in still images via multiple features. In Andrea Fusiello, Vittorio Murino, and Rita Cucchiara, editors, *Computer Vision – ECCV 2012. Workshops and Demonstrations*, pages 263–272. Springer Berlin Heidelberg, Berlin, Heidelberg, **2012**. ISBN 978-3-642-33885-4. doi:10.1007/978-3-642-33885-4_27.

[101] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3686–3693. **2014**. doi:10.1109/CVPR.2014.471.

[102] G. Gkioxari, R. Girshick, and J. Malik. Contextual action recognition with r*cnn. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1080–1088. **2015**. doi:10.1109/ICCV.2015.129.

[103] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587. **2014**. doi:10.1109/CVPR.2014.81.

[104] Rohit Girdhar and Deva Ramanan. Attentional pooling for action recognition. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 34–45. Curran Associates, Inc., **2017**.

[105] Shervin Minaee and Amirali Abdolrashidi. Deep-emotion: Facial expression recognition using attentional convolutional network. *arXiv preprint arXiv:1902.01019*, **2019**.

[106] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164. **2017**.

[107] Hiroshi Fukui, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. Attention branch network: Learning of attention mechanism for visual

explanation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10705–10714. **2019**.

[108] Lu Liu, Robby T. Tan, and Shaodi You. Loss guided activation for action recognition in still images. In C.V. Jawahar, Hongdong Li, Greg Mori, and Konrad Schindler, editors, *Computer Vision – ACCV 2018*, pages 152–167. Springer International Publishing, Cham, **2019**. ISBN 978-3-030-20873-8. doi:10.1007/978-3-030-20873-8_10.

[109] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 17–24. **2010**. doi:10.1109/CVPR.2010.5540235.

[110] Y. Chao, Z. Wang, Y. He, J. Wang, and J. Deng. Hico: A benchmark for recognizing human-object interactions in images. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1017–1025. **2015**. doi:10.1109/ICCV.2015.122.

[111] Yuanjun Xiong, Kai Zhu, Dahua Lin, and X. Tang. Recognize complex events from static images by fusing deep channels. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1600–1609. **2015**. doi:10.1109/CVPR.2015.7298768.

[112] J. Stöttinger, J. R. R. Uijlings, A. K. Pandey, N. Sebe, and F. Giunchiglia. (unseen) event recognition via semantic compositionality. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3061–3068. **2012**. ISSN 1063-6919. doi:10.1109/CVPR.2012.6248037.

[113] L. Bossard, M. Guillaumin, and L. Van. Event recognition in photo collections with a stopwatch hmm. In *2013 IEEE International Conference on Computer Vision*, pages 1193–1200. **2013**. ISSN 1550-5499. doi:10.1109/ICCV.2013.151.

[114] W. Li, Q. Yu, A. Divakaran, and N. Vasconcelos. Dynamic pooling for complex event recognition. In *2013 IEEE International Conference on Computer Vision*, pages 2728–2735. **2013**. ISSN 1550-5499. doi:10.1109/ICCV.2013.339.

[115] Paul Viola and Michael J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, **2004**. ISSN 1573-1405. doi:10.1023/B:VISI.0000013087.49260.fb.

[116] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 487–495. Curran Associates, Inc., **2014**.

[117] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 3320–3328. MIT Press, Cambridge, MA, USA, **2014**.

[118] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826. **2016**.

[119] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. **2015**.

[120] Gokhan Tanisik, Cemil Zalluhoglu, and Nazli Ikizler-Cinbis. Facial descriptors for human interaction recognition in still images. *Pattern Recognition Letters*, 73:44 – 51, **2016**. ISSN 0167-8655. doi:https://doi.org/10.1016/j.patrec.2016.01.002.

[121] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. **2017**.

[122] Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, **1947**. doi:10.1007/BF02295996.

[123] Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y. Ng. Unsupervised learning of hierarchical representations with convolutional deep belief networks. *Commun. ACM*, 54(10):95–103, **2011**. ISSN 0001-0782. doi:10.1145/2001269.2001295.

[124] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for simplicity: The all convolutional net. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*. **2015**.

[125] Daniel Omeiza, Skyler Speakman, Celia Cintas, and Komminist Weldemariam. Smooth grad-cam++: An enhanced inference level visualization technique for deep convolutional neural network models. *ArXiv*, abs/1908.01224, **2019**.

[126] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks, **2019**.

[127] R. Gao, B. Xiong, and K. Grauman. Im2flow: Motion hallucination from static images for action recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5937–5947. **2018**.

[128] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth prediction. **2019**.

[129] H. Fang, S. Xie, Y. Tai, and C. Lu. Rmpe: Regional multi-person pose estimation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2353–2362. **2017**. doi:10.1109/ICCV.2017.256.

[130] Pedro O. Pinheiro, Ronan Collobert, and Piotr Dollár. Learning to segment object candidates. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, page 1990–1998. MIT Press, Cambridge, MA, USA, **2015**.

[131] P. J. Bota, C. Wang, A. L. N. Fred, and H. Plácido Da Silva. A review, current challenges, and future possibilities on emotion recognition using machine learning and physiological signals. *IEEE Access*, 7:140990–141020, **2019**.

[132] J. Guo, Z. Lei, J. Wan, E. Avots, N. Hajarolasvadi, B. Knyazev, A. Kuharenko, J. C. S. Jacques Junior, X. Baró, H. Demirel, S. Escalera, J. Allik, and G. Anbarjafari. Dominant and complementary emotion recognition from still images of faces. *IEEE Access*, 6:26391–26403, **2018**.

[133] Jie Zhou, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, and Maosong Sun. Graph neural networks: A review of methods and applications. *CoRR*, abs/1812.08434, **2018**.

HACETTEPE UNIVERSITY
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING
THESIS/DISSERTATION ORIGINALITY REPORT

HACETTEPE UNIVERSITY
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING
TO THE DEPARTMENT OF COMPUTER ENGINEERING

Date: 18./08/2020.

Thesis Title / Topic: RECOGNIZING HUMAN INTERACTIONS IN STILL IMAGES

According to the originality report obtained by myself/my thesis advisor by using the *Turnitin* plagiarism detection software and by applying the filtering options stated below on 17./08./2020 for the total of ....114..... pages including the a) Title Page, b) Introduction, c) Main Chapters, d) Conclusion sections of my thesis entitled as above, the similarity index of my thesis is ...9..... %.

Filtering options applied:

1. Bibliography/~~Works Cited~~ excluded
2. Quotes excluded / ~~included~~
3. Match size up to 5 words excluded

I declare that I have carefully read Hacettepe University Graduate School of Sciene and Engineering Guidelines for Obtaining and Using Thesis Originality Reports; that according to the maximum similarity index values specified in the Guidelines, my thesis does not include any form of plagiarism; that in any future detection of possible infringement of the regulations I accept all legal responsibility; and that all the information I have provided is correct to the best of my knowledge.

I respectfully submit this for approval.

18.08.2020
Date and Signature

Name Surname: Gökhan TANIŞIK
Student No: N10162705
Department: Computer Engineering
Program: Computer Engineering
Status: ☐ Masters ☐ Ph.D. ☒ Integrated Ph.D.

ADVISOR APPROVAL

APPROVED.

Assoc. Prof. Nazlı İKİZLER CİNBİŞ

(Title, Name Surname, Signature)

CURRICULUM VITAE

Credentials

Name,Surname	:	Gökhan TANIŞIK
Place of Birth	:	Malatya
Marital Status	:	Married
E-mail	:	gokhantanisik@gmail.com

Education

BSc.	:	Computer Engineering, Hacettepe University Ankara, TURKEY, 2009
MSc.	:	-
PhD.	:	Computer Engineering, Hacettepe University Ankara, TURKEY, 2020, (Integrated PhD.)

Foreign Languages

English

Work Experience

Software Specialist, INNOVA , 2009-2010

Software Engineer, ASELSAN, 2010-On Going

Areas of Experiences

Image Processing, Computer Vision, Embedded and Real Time Applications

Publications

- Gokhan Tanisik, Cemil Zalluhoglu, Nazli Ikizler-Cinbis, Multi-stream Pose Convolutional Neural Networks for Human Interaction Recognition in Images, *Computer Vision and Image Understanding (CVIU)*, **Under Review**
- Gokhan Tanisik, Cemil Zalluhoglu, and Nazli Ikizler-Cinbis. Facial descriptors for human interaction recognition in still images. *Pattern Recognition Letters*, 73:44 – 51, **2016**. ISSN 0167-8655.

- Gokhan Tanisik, Erhan Gundogdu, Multiple Model Adaptive Visual Tracking With Correlation Filters, *IEEE International Conference on Image Processing (ICIP)*, Quebec, Canada, September **2015**

- Gokhan Tanisik and Erhan Gundogdu, Bi-Model Visual Tracking with Correlation Filters, *IEEE Signal Processing and Communications Applications Conference (SIU)*, Malatya, Turkey, May **2015**

- Gokhan Tanisik, Oguzhan Guclu and Nazli Ikizler-Cinbis, Bolut ve Kontur Ozniteliklerini Kullanarak Imgelerdeki Insan Hareketlerini Tanima, *IEEE Signal Processing and Communications Applications Conference (SIU)*, Northern Cyprus, April **2013**