TRANSFER LEARNING FOR SIGN LANGUAGE RECOGNITION

by

Ahmet Alp Kındıroğlu

B.S., Computer Science And Engineering, Sabancı University, 2008

M.S., Computer Engineering, Boğaziçi University, 2011

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

Graduate Program in Computer Engineering

Boğaziçi University

2023

ACKNOWLEDGEMENTS

Words cannot express my gratitude to my professor and supervisor of my thesis Prof. Lale Akarun for all the support, guidance and patience she has given me throughout my graduate education. I'm also indebted to my former and current progress committee members Assoc Profs. İnci Baytaş, Kadir Gökgöz and Profs. Sumru Özsoy, Albert Ali Salah; who generously provided their knowledge and expertise throughout my thesis. In addition, I would like to express my deepest gratitude to the defense jury members Prof. Hacer Yalım Keleş and Prof. Hatice Köse. This project also would not have been possible without the assistance of the Tetam Institute and its directors, Profs. Cem Ersoy, Lale Akarun and Tuna Tugcu, who provided me with a scholarship and the computing resources I needed for my research.

I am also grateful that, while I was a Ph.D. student there, the Department of Computer Engineering gave me a vibrant and enjoyable working environment. I would like to thank everyone I have gotten to know and work with throughout my time at the department. I want to thank Berk Gökberk, Suzan Üsküdarlı, Murat Saraçlar, and Oya Aran in particular for their help and advice with my thesis. A special thanks go out to my lab and departmental colleagues Necati Cihan Camgöz, Oğulcan Özdemir, İsmail Arı, Yunus Emre and Gaye Kara, Barş Evrim Demiröz, Ufuk Can Biçici, and Doğa Siyli whose friendship, support, and scientific contributions made my time in PILAB a great experience.

Finally, I'd like to thank my family from the bottom of my heart. I would like to express my deepest gratitude to my spouse Rabia and our son Ege, whose love, encouragement, and everlasting patience over the many years of my thesis have kept my spirits and motivation strong throughout my studies. Furthermore, I'm extremely grateful to my parents Neşe and Ergun and brother Kaan, without whose unwavering support, encouragement, and belief in me I could not have completed this journey.

During this thesis, I was supported by the Scientific and Technological Research Council of Turkey (TUBITAK) with Grant No. 108E113 and 109M101 and Santez project 0341.STZ.2013-2. In addition this thesis was supported by the Turkish Directorate of Strategy and Budget under the TAM Project number 2007K12-873 and Boğaziçi University Research Fund under the Grant Number 16903. The numerical calculations reported in this paper were fully/partially performed at TUBITAK ULAKBIM, High Performance and Grid Computing Center (TRUBA resources). Thanks to these supports, I could concentrate only on my research rather than working outside academia.

The images that emerged within the scope of this thesis work and whose copyrights were transferred to the publishing house were used in the thesis book in accordance with the publication policy of the author on the reuse of the texts and graphics produced by the author on their own web page.

ABSTRACT

TRANSFER LEARNING FOR SIGN LANGUAGE RECOGNITION

Sign languages are visual languages that use hands, arms, and faces to communicate concepts. In the last decade, sign language recognition (SLR) research has made significant progress but still requires massive amounts of data to recognize signs. Despite efforts to create large annotated sign language datasets, applications that can translate for ordinary users in daily settings are yet to be produced. Most SLR research focuses on a few popular sign languages, leaving most sign languages, especially Turkish Sign Language (TID), under-resourced for sign language technology development.

This dissertation addresses several open research questions about the development of SLR technology for TID from several perspectives. We generated BosphorusSign22k, an isolated SLR dataset for TID with 22k videos, and benchmarked state-of-the-art techniques on it. We proposed aligned temporal accumulative features (ATAF) to efficiently model sign language movements as dynamic and static subunits. Combined with methods using other modalities, the method achieves state-of-the-art performance on BosphorusSign22k. We then used regularized regression-based multi-task learning and presented task-aware canonical time warping for isolated SLR. The technique aligned and grouped signs to minimize discrepancies across different sources and emphasize class differences. Finally, we established a benchmark for cross-dataset transfer learning in isolated SLR. We evaluated supervised transfer learning algorithms using a temporal graph convolution-based SLR method. Experiments with closed and partial-set cross-dataset transfer learning reveal a substantial improvement over combined training and fine-tuning-based baseline techniques.

ÖZET

İŞARET DİLİ TANIMA İÇİN ÖĞRENME TRANSFERİ

İşaret dili, kavramları iletmek için elleri, kolları ve yüzleri kullanan görsel bir dildir. Son on yılda, işaret dili tanıma (SLR) araştırmaları önemli ilerleme kaydetmiştir, ancak işaretleri tanımak için hala büyük miktarda veri gerekmektedir. İşaret dilleri için işaretlenmiş büyük işaret dili veri kümeleri oluşturulması için harcanan kaynaklara rağmen, günlük kullanıcılar için çeviri yapabilen uygulamalar henüz üretilebilmiş değildir. Buna ek olarak, işaret dili tanıma araştırmalarının çoğu birkaç popüler işaret diline odaklanmakta ve Türk İşaret Dili (TİD) de dahil olmak üzere işaret dillerinin çoğu gelişen işaret dili teknolojileri için yetersiz kaynaklara sahip diller olarak kalmaktadır.

Bu tez, TİD için işaret dili tanıma teknolojilerinin geliştirilmesiyle ilgili bir dizi açık araştırma sorusunu ele almaktadır. Bu kapsamda, 22 bin video içeren izole bir TİD SLR veri seti olan BosphorusSign22k oluşturulmuş ve bu veri setinde literatürdeki en başarılı yöntemleri kullanarak tanıma sonuçları kıyaslanmıştır. İkinci olarak, işaretleri dinamik ve statik alt birimler olarak modellemek için hizalanmış zamansal birikimli öznitelikler (ATAF) yöntemi önerilmiştir. Geliştirilen yöntem, diğer kipleri kullanan yöntemlerle birleştirildiğinde literatürde BosphorusSign22k veri kümesinde elde edilen en yüksek sonucu elde etmektedir. Ardından, düzenlileştirilmiş regresyona dayalı çoklu görev öğrenme ile SLR yapmak için kanonik zaman hizalama adı verilen, farklı kaynaklar arasındaki tutarsızlıkları en aza indiren ve sınıf farklılıklarını vurgulayan bir işaret hizalama ve guruplama yöntemi önerdik. Son olarak, mevcut iki TİD tanıma veri kümesi kullanarak çapraz veri-kümesi öğrenme transferi için bir kıyaslama veri kümesi oluşturulmuştur. Bu veri kümeleri üzerinde, zamansal çizge sinir ağları tabanlı SLR yöntemleri kullanarak beş denetimli transfer öğrenme algoritması değerlendirilmiş ve dayanak teknikler üzerinde önemli bir gelişme sağladığı ortaya koyulmuştur.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS . . . iii
ABSTRACT . . . v
ÖZET . . . vi
LIST OF FIGURES . . . x
LIST OF TABLES . . . xiii
LIST OF SYMBOLS . . . xviii
LIST OF ACRONYMS/ABBREVIATIONS . . . xix
1. INTRODUCTION . . . 1
2. Literature Review . . . 8
2.1. Sign Languages and Sign Linguistics . . . 8
2.2. Sign Language Recognition . . . 9
2.3. Relevant Frame Selection for Efficient Video Classification . . . 15
2.4. Video Alignment . . . 16
2.5. Transfer Learning . . . 17
3. SIGN LANGUAGE REPRESENTATION AND RECOGNITION . . . 21
3.1. BosphorusSign22k Dataset . . . 21
3.1.1. BosphorusSign22k Phonological Labels . . . 22
3.2. 3D Convolutional Neural Network Based SLR . . . 24
4. LINGUISTIC REPRESENTATIONS FOR SLR . . . 30
4.1. Aligned Temporal Accumulative Features . . . 30
4.1.1. Method . . . 32
4.1.1.1. Keyframe Extraction . . . 32
4.1.1.2. Alignment with Temporal Transformer Networks . . . 34
4.1.1.3. Temporal Accumulative Features . . . 37
4.1.1.4. Classification . . . 40
4.1.2. Experiments . . . 42
4.1.2.1. Analysis of Accumulative Representation Methods . . . 43
4.1.2.2. Ablation Study on Feature Representation Schemes . . . 44

4.1.2.3. Hyper-parameter Optimization for ATAF 47
4.1.2.4. Choosing the Best Set of Joints for TAF 48
4.1.2.5. Using Keyframes to Colorize Temporal Video Segments 49
4.1.2.6. Keyframe Alignment for Better Temporal Modeling . . 50
4.1.2.7. Alignment of Linearly Sampled Frames with TTN . . . 51
4.1.2.8. Fusion Performance in Terms of Sign Attributes 53
4.1.2.9. Comparison with State-Of-The-Art Methods 54
4.1.3. Discussion 57
5. MULTI-TASK LEARNING APPROACHES 60
5.1. Regularized Multitask learning & Personality Prediction 60
5.1.1. Prediction of Personality Traits 62
5.1.2. Data and Annotations 64
5.1.3. Methodology 66
5.1.4. Feature Selection 67
5.1.5. Multi-Task Learning 68
5.1.5.1. Multi-Task LASSO 69
5.1.5.2. $L_{2,1}$ Norm Regularized Multi-Task LASSO 69
5.1.5.3. Trace Norm Regularized Multi Task LASSO 70
5.1.5.4. Robust Multi-Task Learning 70
5.1.5.5. Robust Multi-Task Feature Learning 70
5.1.5.6. Dirty Multi-Task LASSO 71
5.1.5.7. Multi-Task Discriminant Analysis 71
5.1.6. Transfer Learning 72
5.2. Experiments and Results 72
5.2.1. Transfer Learning Framework 73
5.2.2. Multi Task Learning Framework 74
5.2.3. Experiments 75
5.2.3.1. Prediction of Individual Personality Traits 75
5.2.3.2. How Useful Are the Individual Features 76
5.2.3.3. How Beneficial Is Multi-Task Learning 78
5.2.3.4. Can We Transfer Knowledge from a Richer Domain . . 79

5.2.3.5. Is Knowing Too Much Detrimental 80
5.2.3.6. What Is the Best Overall Strategy 82
6. TASK AWARE CANONICAL TIME WARPING 85
6.0.1. Temporal Registration For Sign Language Recognition 86
6.0.1.1. Generalized Canonical Time Warping 88
6.0.1.2. Clustering based Task Aware CTW 89
6.0.1.3. Keyframe Selection 90
6.0.2. Sign Language Recognition 91
6.0.3. Experiments 92
6.0.3.1. Baseline Methods for Classifying Signs 93
6.0.3.2. Temporal Registration of Signs Prior to Classification . 93
6.0.3.3. Negative Transfer Avoidance by Clustering Sequences . 94
6.0.3.4. Analysis of Temporal Strategies for C3D Features . . . 95
6.0.4. Discussion 96
7. DEEP DOMAIN ADAPTATION FOR SIGN RECOGNITION 98
7.1. Technical Approach 99
7.1.1. SL-GCN based Isolated Sign Language Recognition 99
7.1.2. Supervised Transfer Methods 101
7.2. BSgin57 Subset 102
7.3. Experiments and Results 103
7.3.0.1. Baseline Isolated Recognition Results 104
7.3.0.2. Supervised Closed Set Transfer Learning on Bsign22k . 105
7.3.0.3. Partial Set Transfer Learning on Bsign22k 106
8. CONCLUSION 109
REFERENCES 114

LIST OF FIGURES

Figure 3.1. Examples of signing performances from different signers in the BosphorusSign22k dataset. 22

Figure 3.2. An overview of the analyzed sign types. 24

Figure 4.1. Overview of the aligned temporal accumulative features method. The proposed Aligned Temporal Accumulative Features (ATAF) approach uses keyframe detection and keyframe alignment followed by sign language representation and classification. Predictions are then fused with the predictions of 3D-Convolutional Neural Network (CNN)s to improve recognition accuracy. Variants of ATAF without keyframe extraction and alignment are used to validate our proposed approach. 31

Figure 4.2. A visualization of keyframe based alignment. Given sequences of different length belonging to different classes, candidate keyframes from each sign are aligned to create better temporal representations. 34

Figure 4.3. Different keyframe feature representations to calculate alignment. On the left side, hand coordinate based features are processed, while on the right side heatmap-based features are presented. . . . 36

Figure 4.4. Overview of the HSV based coloring method using keyframes. Each keyframe is uniquely color coded with a unique hue channel and a gradient style saturation channel indicating temporal location. The final feature is calculated by stacking joint heatmaps, applying aggregations and converting the resulting image to RGB format. . 37

Figure 4.5. The ATAF colorization for the left hand joints are shown in the left image. In order to represent all joints, similar feature images belonging to all critical joints are given as inputs to the learning model. 38

Figure 4.6. The features U_j (left) and N_j (center) are visualized for a group of hand and body gestures. 40

Figure 4.7. SLR Experiments on Ankara University Turkish Sign Language Dataset (AUTSL) and Bosphorussign22k Datasets. 55

Figure 5.1. Set of personality traits and features used in prediction. VLOG segment displays overlapping set of features and tasks. 66

Figure 5.2. Prediction Accuracy of Transfer Learning in Extraversion Prediction. 80

Figure 5.3. Multi-Task learning using forward selection of tasks. At each step, the best performing personality trait is included in the training set. Prediction performance on ELEA dataset for the leadership trait (left) and the extraversion trait (right). 81

Figure 6.1. Existing indirect alignment approaches vs proposed direct alignment approach. 85

Figure 6.2. Warping sequences with DTW. 87

Figure 6.3. Finding correlated features with CTW. 88

Figure 6.4. Canonical time warping algorithm. 89

Figure 6.5. Task Aware Canonical Time Warping. 90

Figure 7.1. Representative Images from the $BSign22k_{shared}$ (images to the left with green background) and $AUTSL_{shared}$ (images to the right with white & complex backgrounds) datasets. 104

LIST OF TABLES

Table 2.1. A collection of available Isolated Sign Language Recognition Datasets. 11

Table 2.2. Reported state of the art methods on UCF-101 and Kinetics 400 datasets. 12

Table 3.1. Specifications of the BosphorusSign22k (BSign22k) Dataset. 23

Table 3.2. Baseline MC3 ResNet results on the BSign22k dataset. 26

Table 3.3. Classification accuracy results of the sampling and fusion settings. Three different settings are provided in the table. From left to right, (1) Single cue spatial sampling results, (2) Active Window Based Temporal Sampling applied to each crop, and (3) Spatial&Temporal settings are combined in one setting. Note that the bottom two rows include the fusion result of the above three models in each setting. 27

Table 3.4. Effects of excluding individual cue units from the final fusion model. Using the different two cue settings and their performance, we infer to the excluded setting and its effect on the final mix. 27

Table 3.5. Analysis of the temporal sampling based recognition approach with respect to signs with certain grammatical sign attributes: one-handed signs, two-handed signs, mono-morphemic signs, compound signs, and signs involving repetitive and circular movements, respectively. 28

Table 4.1. Effects of the number of linearly sampled temporal subunits parameter on recognition accuracy (%). 44

Table 4.2. Recognition accuracy (%) of different image and kernel sizes. . . . 45

Table 4.3. Ablation study of ATAF with different feature compositions. The displayed results are accuracy (%) values on the first 50 classes of the BSign22k Dataset. 46

Table 4.4. Neural network parameter optimization for Hue based temporal colorization without temporal subunits. 47

Table 4.5. The 25-class accuracy of the linearly sampled ATAF for different joint groups is shown. For each joint group shown as JG# joints belonging to the action group are included in the feature. The S and G columns indicate single joints being added as individual features vs groups of joints in a single feature. 48

Table 4.6. Top-1 accuracies (%) of approaches that use equal length temporal segments (TS) or keyframes detected with *Hand* and *Finger* speeds. The experiments are performed on the first 50, 100 and 744(all) classes of BSign22k dataset. 49

Table 4.7. Comparison of recognition accuracy (%) on keyframe based ATAF effectiveness with and without Temporal Transformer Network (TTN). The experiments are performed on the first 50, 100 and 744(all) classes of BSign22k dataset. 51

Table 4.8. Analysis of the Temporal transformer based method without using keyframe selection. Linear sampling with different number of frames (# of frames) are used to select (ts) number of temporal segments. 52

Table 4.9. The accuracies (%) of Linear sampling, the keyframe-based method, TTN Alignment of Keyframes and TTN alignment of linearly sampled frames are reported on different gesture subgroups: One-handed (1H), two-handed (2H), circular (C), non-circular (NC), repetitive (R), non-repetitive (NR), monomorphemic (MM), and compound (COM) signs. 53

Table 4.10. The accuracies (%) of score fusion of the Conv3D MC3 method with Linear sampling, the keyframe-based method, TTN Alignment of Keyframes and TTN alignment of linear sampled frames are reported on different gesture subgroups: One-handed (1H), two-handed (2H), circular (C), non-circular (NC), repetitive (R), non-repetitive (NR), monomorphemic (MM), and compound (COM) signs. 54

Table 4.11. Comparison of ATAF using coordinate (ATAF-CD-KF) and heatmap (ATAF-HM-KF) based keyframe detection approaches with state-of-the-art heatmap based SSTCN and Motion History Image with RGB fusion and coordinate based Hierarchical Co-occurrence Network (HCN) and Sign language graph convolution networks (SL-GCN) methods. 56

Table 4.12. Analysis of clip length for the comparison of ATAF using coordinate-based keyframe(ATAF-KF), linear sampling(ATAF-LINEAR SAMPLING), and heatmap based keyframe detection (ATAF-HM-KF) approaches with state-of-the-art heatmap based SSTCN, HCN and GCN methods. 57

Table 5.1. List of non-verbal features extracted from the ELEA dataset. . . . 65

Table 5.2. List of Classifiers used in the study. 75

Table 5.3. Single task prediction accuracies for personality traits. 76

Table 5.4. Prediction performance for individual feature groups. 77

Table 5.5. Effects of feature selection and extraction approaches for single task prediction for extraversion and leadership. 78

Table 5.6. Multi-task recognition for leadership and extraversion traits by utilizing all personality and leadership traits on ELEA. 79

Table 5.7. Performance on the ELEA Dataset. 83

Table 6.1. Baseline classifier accuracies for BS-SSS Dataset. 93

Table 6.2. Analysis of the alignment methods used for temporal registration. . 94

Table 6.3. Exploring the effect of task wise clustering on canonical time warping. 94

Table 6.4. Exploring combined the effect of user wise grouping combined with task wise clustering on canonical time warping. 95

Table 6.5. Evaluation of SLR accuracy with linearly spaced samples temporal strategy. 96

Table 6.6. Evaluation of SLR accuracy with random samples from linearly spaced intervals strategy. 96

Table 6.7. Evaluation of SLR accuracy with keyframe based linearly spaced intervals strategy. 97

Table 7.1. Number of identical sign pairs and total number of videos from the AUTSL and Bsign22k datasets 103

Table 7.2. Baseline accuracy results for isolated sign language recognition when no prior information on shared classes to be transferred is present. 105

Table 7.3. Top-1 Accuracy results for closed set transfer learning are reported. Each method uses $\text{AUTSL}_{\text{shared}}$ as source and $\text{BSign22k}_{\text{shared}}$ as target domains. 106

Table 7.4. Top-1 Accuracy results for partial-set transfer learning are reported. Each method uses AUTSL as source and $\text{BSign22k}_{\text{shared}}$ as target domains. 107

Table 7.5. Top-1 Accuracy results for fusion of transfer learning methods. Each method uses AUTSL as source and $\text{BSign22k}_{\text{shared}}$ as target domains. 108

LIST OF SYMBOLS

$d_{n,j}$	Local Minima values for displacement
f_c	Cutoff threshold
f_{P_s,P_k}	Point distance between two points
h	Height of image
H_j	Hue Image.
I	Sign language video as a collection of frames
K	Number of selected Keyframes
M	Number of detected candidate Keyframes
N_j	Normalized ATAF Feature.
$P_{m,j}$	Speed of each candidate keyframe m for joint j
S_j	Saturation Image.
T	Number of frames in a video.
U_j	Unnormalized ATAF Feature.
V_j	Accumulated Heatmap Image.
w	Width of image
$X_{D,T}$	Numerical coordinates representing the sign in every frame
$\gamma(t)$	Value of monotonically increasing function at time step t
$\Gamma(T)$	Index function for aligning a sequence
$\delta s, j$	Local density
η	Alignment function
ϕ_s	Distance from the closest point

LIST OF ACRONYMS/ABBREVIATIONS

3DCNN	3D Convolutional Neural Network
ADDA	Adversarial Discriminative Domain Adaptation
ASL	American Sign Language
ATAF	Aligned Temporal Accumulative Features
AUTSL	Ankara University Turkish Sign Language Dataset
BSign22k	BosphorusSign22k
BSL	British Sign Language
CNN	Convolutional Neural Network
DANN	Domain Adversarial Neural Network
DGS	German Sign Language
DSBN	Domain Specific Batch Normalization
HMM	Hidden Markov Model
HCN	Hierarchical Co-occurrence Networks
HSV	Hue Saturation Value
HAR	Human Action Recognition
JAN	Joint Adaptation Network
MDD	Margin Disparity Discrepancy
MCC	Minimum Class Confusion
MHI	Motion History Images
MRMR	Minimum Redundancy Maximum Relevance
Potion	Pose Motion Representation for Action Recognition
PCA	Principal Component Analysis
SL	Sign language
SLR	Sign language recognition
SSTCN	Separable Spatial Temporal Convolutional Networks
TID	Turkish Sign Language
TTN	Temporal Transformer Network

1. INTRODUCTION

Sign languages are the native languages of deaf communities. They are visual languages where users convey meaning using configurations and movements of hands, upper body and face. Deaf people use sign languages on a daily basis to communicate with each other and with hearing people. Along with complementary communication skills like cued speech, lip reading and communicative gestures, sign languages are an effective mode of communication that can only be used when both parties of a conversation are fluent in its vocabulary and grammatical rules. As studies show that literacy among the deaf population is lower than the rest of society [1], the participation and communication of deaf people in public life are limited by a communication barrier. To establish communication with computers or users who are not fluent in sign language, designing systems that recognize and translate sign language into a spoken language or the written text of a spoken language would be a very useful innovation. Researchers of computer vision and pattern recognition has been working towards making such applications feasible for the last forty years with great progress [2–4]. Unfortunately, compared to more popular linguistic pattern recognition technologies such as speech recognition or text recognition, sign language recognition research has not yet produced reliable applications that can perform translations for ordinary users in daily tasks.

The field of sign language recognition research is a multidisciplinary field including research elements from pattern recognition, computer vision, linguistics and natural language processing communities. Signs in sign languages are composed of shape and movement of hands, upper body movements and facial gestures, which form their manual and non-manual components. Sign language recognition (SLR) aims to recognize these signs and map them to glosses, which are the spoken language correspondences/transcriptions of these signs. Glosses correspond to the smallest linguistically meaningful components in sign languages, similar to simple words in spoken languages. Translation of a sign language video requires the visual capture of manual and non-manual components from the signing space, fitting a linguistic grammar model

to its manual and non-manual parts and using that model to achieve translation to a spoken language. In this study, we are interested in isolated SLR, which aims to solve the computer vision sub-problem of accurately capturing and recognizing glosses.

The task of achieving effective isolated SLR is an active research problem with several challenges. First of all, there exist many sign languages, which adds an additional language challenge to SLR research. According to Woll *et al.*, there exist more than 103 sign languages that differ significantly from one another in both phonology, morphology, syntax and grammar. These languages also differ from the spoken languages of the surrounding majority community. For example, written Turkish and Turkish Sign Language (TID) differ in vocabulary, word order, phonology and syntactic rules. This creates challenges to data collection and annotation, which is vital for successful pattern recognition research and development in SLR.

For sign languages, annotating video is a time-consuming task. Expert signers annotate glosses in the order they appear in the video. This difficult task consumes 10-30 minutes for a minute-long sign video using linguistic annotation tools such as ELAN [5]. While large scale annotation efforts exist for some languages such as DGS [6], BSL [7], ASL [8], most of the remaining languages are still considered under-resourced languages. In addition to the problems stemming from data collection or annotation, maintaining standardization and high variance in both syntax, vocabulary and grammar also cause problems for SLR. On top of this, in many existing datasets, capturing data from the 3D signing space using 2d images often has a detrimental effect on the performance of SLR. Imperfect capture and representation of human joints such as fingers make ignoring variances between different data sources harder than distinguishing different glosses from one another [9].

For this reason, to train high-quality isolated SLR models under such constraints, we hypothesize that striving to utilize different sources of data in training isolated SLR models may improve recognition performance for under-resourced languages like TID. The motivation of this thesis is to produce an isolated sign language recognition

methodology that can be used to effectively recognize sign language gestures using annotated data available from different sources. The most commonly used approach is to utilize transfer learning from similar datasets or tasks to improve recognition accuracy. To achieve this goal, we focus on two open research problems in isolated sign language recognition and combine divergent research paths in these two fields to prove our hypothesis. These are the development of a baseline sign language recognition module that can reflect the linguistic characteristics of sign languages and transfer learning techniques, which employ data from sources with lots of data to enhance performance on tasks with little data.

In chapter two of this thesis, we give a detailed literature survey on the current state of the art in sign language recognition and transfer learning. In chapter three, we elaborate on our efforts to establish a baseline isolated SLR methodology to recognize glosses from a large vocabulary of signs. We generated BosphorusSign22k [10], an isolated SLR dataset for TID containing 22k videos from 744 different classes and provided benchmark results using state-of-the-art approaches on this dataset.

Chapter four focuses on finding efficient representations of sign languages that can effectively model signs based on their phonological properties. Isolated SLR is a video classification problem where the task is to assign the correct sign gloss label to a video. For this reason, research in this field closely followed action recognition, which constructed huge datasets from thousands to millions of videos, and used Spatio-temporal architectures to model human behavior in videos. Such models used different modalities of data such as RGB [11], depth [12] and optical flow images [13] as well as detected human joints [14] to model human actions using different state-of-the-art techniques such as HMM's [15], 3D-CNN's [13], graph neural networks [16] or Vision Transformers [17]. However, ideally being designed for the action recognition domain where context independent of a user is just as important, the information in sign language recognition lies solely in the user's configuration of hands, body and face. For this reason, we designed a representation model for sign language recognition (slr) that would be capable of modeling sign language signs considering the problems' unique

characteristics. We proposed ATAF descriptors capable of accumulating joint coordinates over time to create fixed-sized compact representations from variable-length sign language frame sequences. These features reduced redundant background information in the sign language representation to a minimum. In addition, we utilize multiple configurations of moving and static joints to infer grammatical keyframe information and created a robust baseline classifier tailored to sign language recognition. Finally, we align the detected keyframes in an end-to-end fashion during recognition to construct robust classifiers capable of finding distinguishable alignments for distinct keyframes to maximize recognition performance. Results were presented on the BosphorusSign Dataset, which we collected for a project that has become a part of this thesis.

In chapters five and six, we detail our work on transfer learning and how we came to adapt our work on spatial transfer learning to the problem of sign language alignment. We focused on regularized Multitask learning (MTL) based approaches for this task. Image sequences containing articulations of the same sign show variety due to phonetic environments, singer differences and capture methods. Consequently, modeling signs with various qualities, such as length, form, or signer, can be beneficial in recognition. Learning several related tasks together can improve modeling accuracy and extracted features, especially when each task has a small amount of training samples. When modeling machine learning or transfer learning problems, tasks are not always correlated or related. For this reason, there is a benefit in representing task relationships with multi-task learning approaches. We chose personality prediction as a toy problem to implement and develop regularized regression-based methods on a spatial multi-modal prediction framework prior to moving our experiments to the temporal sign language videos. Several regularized regression-based multitask learning strategies were investigated [18, 19]. Taking popular feature extraction/selection approaches such as Principal Component Analysis (PCA) [20] and Minimum Redundancy Maximum Relevance (MRMR) [21] for preprocessing and classifiers such as SVMs, Ridge regression and decision forests, we compared our transfer and multitask learning approaches to classical machine learning approaches. In the next step, we moved our experiments to the sign language recognition problem by using multi-set alignment [22, 23]

approaches to align sign language gestures and proposed a method called multitask CTW by combining the Canonical Time Warping framework with the spatial domain regularized multitask learning approaches.

In chapter seven of this thesis, we explored the use of supervised transfer learning with deep neural networks to build a SLR methodology capable of maximizing transfer learning accuracy for under-resourced sign languages such as TID. Supervised Transfer learning approaches aim to use data from data-rich sources to improve task performance on a data-poor target task. For isolated SLR, our focus is on TID, where two medium-scale public datasets exist. These are the BSign22k [10] dataset with 22542 videos for 745 distinct signs and the AUTSL [24] dataset with 32k videos for 216 signs. Compared to common video classification datasets such as Kinetics [25], or MomentsInTime [26] with thousands of samples per class, these datasets, along with most other isolated SLR datasets, can be considered as under-resourced datasets.

It is reported that transferring knowledge from larger datasets often improves accuracy, while little benefit is observed when a transfer is attempted between two similarly sized datasets [16, 27]. Although transfer learning datasets exist for video classification [28], there was no SL transfer dataset to test different transfer strategies. For this reason, we created a sign language recognition baseline for cross-dataset sign language recognition using two public isolated Turkish Sign Language datasets. These datasets share 57 common gestures. However, not all gestures are identical, as their performance varies from dataset to dataset due to differences in interpretation and style.

Using this dataset, we experimented with three different transfer learning settings. We construct a transfer learning pipeline using the SL-GCN algorithm as our baseline feature extractor. Existing transfer learning methods make use of combined training and finetuning based methods [16, 27]. In contrast, to transfer signs between two supervised source domains, we use five different transfer learning techniques, namely, finetuning, Domain Adversarial Neural Network (DANN), Minimum Class Confusion

(MCC), Joint Adaptation Network (JAN) and Domain Specific Batch Normalization (DSBN). In setting 1, common signs between source and target datasets are unknown. In setting 2, a closed set transfer learning problem is observed where source and target class labels are identical and known. Setting 3 is a partial set transfer learning problem where the target set contains a subset of class labels belonging to the source set. In all three settings, we tried to find transfer learning approaches that yield the largest improvement compared to baseline approaches.

The outputs and contributions of our work on sign language recognition and transfer learning domains can be listed as follows:

- We collected a large-scale Sign Language Recognition dataset called BSign22k and made the dataset publicly available for research. At the time it was collected, the dataset was the largest TID dataset composed of sequences of RGB images, depth images and human body coordinates for SLR [29, 30]. We performed benchmark SLR experiments on BSign22k with state-of-the-art methods and presented baseline recognition results.
- We proposed a novel temporal sign language representation inspired by sign language linguistics, which can model a gesture with an arbitrary number of static subunits and moving parts using a distinct coloring scheme to represent keyframe ordering [31].
  (i) We proposed a static subunit detection method that uses local minima in hand speed with density clustering to extract dynamic and static sign language subunits.
  (ii) We adopted a temporal transformer-based end-to-end alignment approach to align static and dynamic keyframes and linearly sampled gesture frames to allow better temporal modeling [32].
  (iii) We performed score fusion with 3D-CNN approaches using different modalities to improve isolated SLR performance.
- We proposed a method called task-aware canonical time warping to adapt multi-task regularized regression framework to the problem of sign language alignment.

In the experiment, we compare the results of SLR with different alignment strategies to demonstrate the effectiveness of grouping sign language gestures according to shared features together to maximize recognition performance.

- We proposed a transfer learning methodology to perform SLR in under-resourced sign languages.
  (i) We proposed a novel sign language recognition baseline for cross-dataset sign language recognition using two public isolated Turkish Sign Language datasets.
  (ii) Using the SL-GCN algorithm as our baseline feature extractor and five different transfer learning techniques, namely, finetuning, DANN, MCC, JAN and DSBN, we recognize gestures in three different supervised-transfer learning settings [32].

Finally, in chapter eight, we present the conclusions of the thesis, reaffirm our contributions and present our final takeaways. We discuss current concepts and potential paths for future sign language research.

2. Literature Review

This chapter describes the most relevant literature on sign language recognition and transfer learning. We introduce the reader to essential sign language linguistics before moving on to sign language recognition and transfer learning methods. For computational sign language analysis, we examine the literature on representing videos with fewer frames and video alignment challenges in addition to isolated sign language recognition, in line with the focus of the contributions of this thesis.

2.1. Sign Languages and Sign Linguistics

Our research problem in SLR involves the task of capturing different components of visual actions from the signer by hands, arms, face and torso. Sign languages, having developed independently of spoken languages, have unique rules and grammar that differentiate how different concepts are represented. In sign languages, a sign is composed of handshapes, movement and location according to the Stokoe grammatical model [33, 34]. Over time, different researchers have proposed different models concerning how these components interact with each other. Battison *et al.* [35] added orientation to the list of structural sign components, while Liddell *et al.* [36] introduced the concept that signs are composed of sequences of moving and static parts. The Liddell Johnson grammatical model defined signs as consisting of sequences of movements and holds. These atomic units can be considered as temporal sub-units of sign languages and are called phonemes or signemes in different studies. Most works in sign language linguistics following the works of [33, 36] accept and build on these models' foundations.

In sign linguistics, the movements, orientation and appearance of hands and arms are called the manual components of sign language. Manual components of the signs can be one or two-handed gestures. The signer performs one-handed gestures with their dominant hand, which may or may not change according to their left or right-

handedness. Signs in sign language can be classified as monomorphemic, consisting of a single morpheme or compound/complex signs which are composed of more than one morphemes.

In addition, there are many constructs of sign language that need to be addressed independently [37]. Such constructs and grammatical rules make the problem of sign language recognition much more complex compared to problems such as hand shape or gesture recognition, where the main objective is to distinguish handshapes from one another.

When distinguishing signs for recognition purposes, glosses and classifiers are the two most common types of gestures. While glosses are signs with meaning and can be mapped to spoken language concepts, classifiers describe information such as movement, location, shape or interaction of previously described items. Non manual features such as facial expressions using eyes, eyebrows and mouth can modify the meaning of a sign by turning it into a negation or question. Placement or indexing is when a signer refers to a person or object and assigns it to a place in the signing space. Finally, fingerspelling is used to spell a word in a spoken language with fingers when the signer needs to communicate a word that does not have a mutually known sign by the communicating parties.

2.2. Sign Language Recognition

In the literature, there have been numerous studies working on sign language recognition from images and image sequences for nearly 35 years [2,38–40]. Since then, as larger resources were curated, the technologies used for capture and processing improved and new algorithms for capturing and recognizing signs developed. Due to the inherent challenges in sign language recognition, research in SLR focuses on three problems which are different abstractions of the same problem. These are isolated sign language recognition, continuous sign language recognition and sign language translation [17].

Sign language translation aims to achieve direct translation between spoken and sign languages. Since the syntax and morphology of the languages are different, the translation model relies on NLP-based language modeling. That requires huge datasets and powerful models capable of learning sign language grammar from that data. As large sign language datasets large enough for accurate translation are still newly being made available [6, 7, 41] at the composition of this thesis, researchers have opted to focus on the continuous SLR problem where the goal is to recognize a sequence of signs and match them to their respective glosses. In continuous SLR, sign spotting and sign recognition tasks are performed simultaneously. Of these two tasks, isolated SLR narrows the scope of the problem further by focusing on spatiotemporal sign recognition from image sequences. As the beginning and end time frames of glosses are provided to the model, the sign-spotting step is omitted. In isolated recognition, the focus is on successfully capturing gestures and modeling them in order to distinguish a sign from a large vocabulary of possible other signs. Since most isolated SLR datasets contain gestures that begin and end with a rest pose where hands are lowered, the effect of sign transitions effecting signs called co-articulation is usually a smaller problem compared to continuous SLR. Instead, resolving problems, such as creating better spatiotemporal models that can better represent signs taking into differences between signs and signer performances, can take precedence.

Isolated sign language recognition is a video classification problem. Following the success of deep learning approaches on tasks such as image classification, object detection and segmentation [42–45], recent studies on computer vision-based human behavior understanding have focused on adapting successful deep image recognition architectures into the video classification domain. In domains such as action recognition, video datasets with thousands of samples per class like Kinetics [25], or MomentsInTime [26] allow researchers to train powerful models capable of training temporal models with billions of parameters. While not on the same scale in terms of size, in the last 15 years, many isolated SLR datasets were made publicly available that drove research in Sign Language recognition forward. A selection of these datasets can be seen in Table 2.1.

Table 2.1. A collection of available Isolated Sign Language Recognition Datasets.

Reference	Year	Dataset	Sign Language	Videos #	Signers	Sample/Sign	Classes #	Isolated	Modality
[46]	2007	GSL	Greek (GSL)	840	6	42.0	20	Yes	RGB
[47]	2011	ASLLVD	American (ASL)	9794	2742	3.6	2742	Yes	RGB
[48]	2012	DGS Kinect 40	German (DGS)	3000	15	75.0	40	Yes	RGB, Depth
[49]	2013	IsoGD	Spanish (LSE)	47933	21	192.5	249	Yes	RGB, Depth
[50]	2014	DEVISIGN	Chinese (CSL)	24000	8	12.0	2000	Yes	RGB, Depth
[51]	2015	PSL Kinect 30	Polish (PSL)	300	1	10.0	30	Yes	RGB, Depth
[52]	2016	LSA64	Argentina	3200	10	50.0	64	Yes	RGB, Depth
[53]	2016	LSE-Sign	Spanish (LSE)	2400	2	1.0	2400	Yes	RGB, Depth
[54]	2018	CSL-500	Chinese (CSL)	125000	50	250.0	500	Yes	RGB, Depth
[30]	2018	BosphorusSign22k	Turkish (TID)	22542	6	30.3	744	Yes	RGB, Depth
[41]	2018	PHOENIX14T	German (DGS)	67781	9	63.6	1066	No	RGB, Depth
[55]	2018	MS-ASL	American (ASL)	25513	222	25.5	1000	Yes	RGB, Depth
[56]	2018	SMILE	Swiss (DSGS)	9000	2281	90	100	Yes	RGB, Pose
[57]	2019	WL-ASL	American (ASL)	21083	119	10.5	2000	Yes	RGB, Depth
[58]	2020	AUTSL	Turkish (TID)	38336	43	169.6	226	Yes	RGB, Depth
[59]	2020	Include	Indian (ISL)	452K	39	198	2281	Yes	RGB, Pose
[60]	2021	BOBSL	British (BSL)	4000	7	16	263	Yes	RGB, Pose

In order to capture manual and non-manual sign performances from image sequences, SLR methods in the literature use a variety of different modalities. While earlier methods used methods such as gloved instruments [61], colored gloves [62] or skin color filtering-based methods [63] to capture different signer articulators, these methods became mostly obsolete with the advent of machine learning based coordinate detection methods [64–66]. Recent works on isolated sign language recognition use different modalities such as RGB images, depth images, optical flow images or human joint coordinates extracted from depth or RGB images.

In terms of video classification methods, models developed for the action recognition domain with large datasets were often successfully transferred and employed in the SLR domain with success. For this reason, exploring the development of video classification in both domains is logical. A chronological listing of state of the art works in video classification methods is given in Table 2.2.

Table 2.2. Reported state of the art methods on UCF-101 and Kinetics 400 datasets.

Method	**Author**	**Year**	**UCF-101**	**Kinetics 400**
IDT	Wang et al.	2011	87.9	-
DeepVideo	Karpathy et al.	2014	63.3	-
Two-Stream Networks	Simonyan et al.	2014	88	-
C3D	Tran et al.	2015	85.2	-
LSTM	Ng. et al.	2015	88.6	
LTC	Varol et al.	2016	92.7	-
TSN	Wang et al.	2016	-	73.9
I3D	Carreira et al.	2017	-	72.1
R2+1D	Tran et al.	2018	-	74.3
Slowfast	Feichtenhofer et al.	2019	-	79.8
TSM	Lin et al.	2019	-	74
CSN	Tran et al.	2019	-	79.1
X3D	Feichtenhofer et al.	2020	-	80.4
TPN	Yang et al.	2020	-	78.9
Vit	Dosovitskiy et al.	2020	-	78.6
Mvit	Fan et al.	2021	-	81.2
Swin-L	Liu et al.	2021	-	84.9

Earlier work on action recognition mostly used constructed features such as improved dense trajectories [67] or 2D CNNs to extend the spatial domain temporally by using different input frames and layer-level fusion techniques [68], and two-stream architectures [69]. In order to improve modeling of videos in the video classification domain, researchers focused on 3 different aspects namely accuracy, efficiency and model generalization. In order to increase model accuracy and generalization, methods for modeling temporal data such as LSTM's [70] or Long-term temporal convolutions [71] were utilized. Later approaches tried to learn videos by exploiting the spatial domain into the temporal domain using 3D CNNs [72] with great success. When performances on smaller datasets such as UCF-101 surpassed 90%, researchers moved on to larger datasets such as Kinetics-400 [73], which simultaneously allowed training better models and provided a much tougher benchmark. Combined solutions utilizing techniques such as multi-stream I3D [73] and R2+1D [74] networks became the sota models both in video and sign language recognition. However, at this stage, efficiency in model design became the area where improvement became possible. Methods with better mechanisms for handling spatial redundancies in data such as slowfast netorks [75], TSM [76] and CSN [77] showed good performances in the action recognition domain. Likewise following the success of scalable models like Efficientnet [78], X3D [79] model also provided an efficient way to handle the tradeoff between model parameter size and complexity. Recurrent models were also still used for working with longer image sequences [80–83]. Finally, the adaptation of vision transformers such as Swin transformers [84] to video recognition domain again made it possible to increase model accuracy and generalization further, pushing state the art results on many benchmark video classification datasets.

On top of models working on the rgb and depth channels, multi-modal methods making use of RGB and flow information simultaneously became popular representations for problems where simultaneous modeling of spatial and temporal information was beneficial. Based on the two-stream hypothesis, which suggests that the human visual cortex consists of two distinct pathways, dorsal (motion) and ventral (appearance) streams, [69] used still frames for ventral and stacked optical flow images for

dorsal streams, and successfully trained both networks on the UCF-101 and HMDB datasets.

Although the idea of two-stream networks was thought of as networks trained on two classification streams (appearance and motion) [69], researchers have expanded this idea by adding multiple modalities (e.g., audio spectrograms) for multi-modal action recognition [85]. The I3D architecture [73] was proposed as an improvement for two-stream networks. In another work, Wang et al. [86] have proposed Temporal Segment Networks (TSNs) with the purpose of solving the long-range temporal limitations of two-stream networks by using temporal sampling. In SLR using RGB video frames directly, state-of-the-art approaches use 3D-CNNs such as I3D [55, 87], coupled with Hidden Markov Models (HMMs) and LSTMs [88] and transformers [17].

Another popular modality that became popular in both action and sign language recognition was coordinate detection-based approaches. Beginning with the work of [64] that became widely available with the Kinect camera, the collection of action and sign language recognition datasets with joint human information became feasible. The downside of these datasets was that it was not possible to extract coordinates from existing video sources that were not recorded with depth cameras. This changed with popular coordinate extraction frameworks such as OpenPose [65], MMpose [89], MediaPipe Holistic [66] libraries. There exists a large number of studies using human joints to perform video classification. Two different representation approaches considered using skeleton joints as numerical coordinates [16, 90] or heatmap images [16, 91].

In recent years, models that model coordinates directly have recently become popular. [55] uses Hierarchical Co-occurrence networks to model gestures, [92] uses transformers, while several other works use Graph Convolutional Networks(GCN). One GCN-based method that is currently state-of-the-art in SLR is SL-GCN [16, 93]. It is based on the Spatio-temporal graph convolutional networks (ST-GCN) method from the action recognition domain, which combines temporal convolutional neural networks with spatial graph-based convolutions of joints [94]. A method called drop-graph [95]

enhances the ST-GCN method by using an attention-based node drop mechanism called drop-graph for regularization, which is then used in isolated SLR by SL-GCN [16, 93] by further adding spatial and temporal attention mechanisms to each layer.

In skeleton heatmap-based recognition approaches, input images are created from numerical coordinates to leverage 2D/3D convolution operations. One technique that utilizes these heatmaps is Separable Spatial Temporal Convolutional Networks (SSTCN) [16] where temporal convolutions are used with 2D convolution operations separately on temporal and spatial channels. Other techniques that use heatmaps are Pose Motion Representation for Action Recognition (Potion) [14], Dynamic Motion Representation [96], and [91] which accumulate joint heatmaps over time and use representation methods such as temporal aggregation [14], or CNN encoders [96] to obtain fixed-size feature vectors from a variable-length video snippet.

2.3. Relevant Frame Selection for Efficient Video Classification

Representing videos with fewer relevant frames has attracted attention in different domains. Examples include video summarization [97], shot boundary detection [98], keyframe detection [99], keyshot detection [100], and action localization [101]. As a means of data abstraction and weak temporal localization, keyframe extraction techniques have been used in the SLR literature over the years [102].

Earlier work on this topic often focused on content changes by selecting keyframes or keyshots, which can significantly reduce the amount of data processed and the complexity of the model without affecting recognition performance. In [103], keyframes are selected by their discriminative power. Carlsson and Sullivan [104] identify distinct activities in a long video clip by matching frame shape information to keyframe prototypes. In [105], a change point model is used to detect significantly different frames in a stream of videos. In [106], density clustering is employed with grayscale image entropy to calculate peaks in entropy change. Density clustering is a method that uses the local density of feature points to identify regional groups [107].

In a similar vein, the use of atomic sign language elements to model sign videos has been intensively investigated in the isolated SLR domain. In contrast, earlier clustering methods solely rely on the distance between feature points; hence they are unable to discover local clusters. Similar to the representation of utterances by phonemes in spoken languages, signs may be considered to be composed of subunits concatenated over time. There are studies on the sequential structure of sign languages, such as the Stokoe notation [33] and the works of Liddell and Johnson [36]. In applying phonetic modeling to automatic recognition, there are several studies. As there is a lack of subunit corpus and lexicons available for most sign languages, the majority of the methods focus on unsupervised learning from data. Some of the notable ones include [108], where the authors use single states in an HMM to cluster dynamic subunits and generate a subunit lexicon. In [109], a rule-based phoneme extraction method is used. One study that makes use of the movement-hold model to construct an end-to-end sign language recognition model is [110]. In other works such as [110] or [111], certain lexical properties such as hand speeds, hand location and hand shapes were used to construct linguistically meaningful subunit lexicons. De *et al.* [112] used architectures that benefited intermediate short temporal subunits to improve model recognition performance. Other studies, such as WLASL-LEX [113], focus on recognizing phonological properties of signs which they further hope to use in the recognition of sign languages.

2.4. Video Alignment

Video alignment for matching two or more videos is also a highly studied research problem. The most commonly used method is Dynamic Time Warping (DTW) [114]. The original DTW algorithm has several drawbacks. It is not differentiable and not invariant to the order of samples. Soft-DTW [115] offers a differentiable loss function that can be used as a layer for end to end alignment with neural networks. DTW Barycenter Averaging [116] is a global averaging method that yields the average of several sequences that is invariant to the ordering of the input sequences. Several studies extract correlated features from multimodal data and align them iteratively with Canonical Time Warping (CTW) [22]. Deep learning versions of the CTW [117]

approach have been shown to work well with noisy samples.

Moreover, researchers have also explored alignment-based frameworks by extending CNN architectures. In [118], a few shot video classification frameworks are constructed via temporal alignment of human motions. A similar method is the SCSampler [119] where a "clip-sampling" model is used to spot clips containing meaningful information from longer videos.

Lohit et al. [120] proposed Temporal Transformer Networks (TTNs) for sequence alignment. The paper proposes a warping function that transforms an input sequence before learning the warped feature using a classification network. The resulting architecture performs discriminative alignment and classification, intending to maximize classification performance. A similar study by Oh et al. [121] also uses a discriminative warping-based approach for a linear time scale.

2.5. Transfer Learning

Supervised machine learning systems work by modeling data from an annotated dataset and applying that model to a validation or test set. When there is abundant data, and the distribution of the data or labels is not divergent from that of the validation set, machine learning models show high performance. However, in real life, this is not always the case, as annotated data may not always be available, or the distribution of the training data might be significantly different from the target test set. In such situations, transfer learning aims to utilize a larger dataset or model from a different source to transfer knowledge and improve performance on a target task. Transfer learning is a very large and active field of research that we can categorize depending on the differences in the data and tasks we wish to transfer. A survey of the recent methods in the field is provided by Wang *et al.* [122].

In the most common supervised transfer learning setting, we can define source data and its labels as x_s, y_s and target data and its labels as x_t, y_t. For problems where x_s and x_t have the same features and labels, but the features or labels come from different distributions, supervised transfer learning methods aim to maximize the recognition performance of the trained model on the target dataset [123]. When the labels belong to similar or different tasks, semi-supervised [124] or unsupervised transfer [125] learning methods are necessary to utilize target data in training.

In the supervised transfer learning literature, different approaches have been proposed over the years to utilize successful model transfer. Prior to deep learning based approaches, instance reweighing [126], feature importance-weighting [127], subspace mapping [128] and regularization based approaches [18] were popular. With the increase in popularity of artificial neural networks finetuning became the most commonly used finetuning [129] approach. Other approaches such as use of source-specific augmentation [130], model weight sharing layers [131], distribution metrics such as mmd [132], consistency regularization [133], adversarial methods [134] and generative methods are examples of recent popular transfer learning methods.

One subfield of transfer learning we focus on in this thesis is multi-task learning, where more than two sources are used for the transfer, and we are interested in maximizing performance in one or more tasks. Multi-task learning methods in the literature work in a supervised classification or regression framework where models for similar tasks are learned simultaneously. These methods focus on improving the overall prediction performance by sharing information between tasks. Multi-task learning-based approaches assume that similarities and differences exist among different tasks, and they employ two different approaches to jointly model them. Earlier work on the topic makes use of a lasso-like regularized regression framework. Evgeniou et al. [135] first made use of regularized regression concerned with capturing these shared structures among tasks.

Following the regularized regression approach, many different models with different regularization methods were proposed. Among these methods, Argyriou et al. [136] tried to find a shared low-dimensional representation for all tasks using L_{21} norm regularization and used these shared features to find task-specific learners for each class. In [137], Gong et al. propose using the L1 norm to find outlier tasks while simultaneously using it on rows to find shared features. They then remove the outlier tasks and perform L_{21} norm multi-task learning on the clean dataset composed of similar tasks. In Jalali et al.'s work [138], the sum of two matrices is used to represent the parameters, and these are regularized differently to learn both shared features and individual outliers for different tasks separately. Follow-up works extended and generalized these concepts by focusing on issues such as learning the relatedness of tasks. Some examples include the work of Kang et al. [139], who used a clustering-based approach with the MTL regularization method to group similar tasks together, and Chen et al., who used a graph-based structured regularization approach to encode the similarity information between tasks [140].

Compared to image-based domain adaptation (DA), Video-based DA attracted fewer studies. In video classification, models have to take into account the temporal variations on top of variations in the image space. In the domain of video domain adaptation, until recently, few studies existed on small-scale datasets such as [28] and [141]. These works focused on the few number of overlapping classes among larger video datasets. Several methods aim to use image-based domain adaptation methods with 3D classification networks, such as adding a gradient reversal layer for domain invariance [142]. The TA3N study [28] proposes a method using domain-specific attention while learning to align frames across domains. They demonstrate the improvement in performance by introducing the UCF-HMDB$_{\text{full}}$ and Kinetics-Gameplay datasets, providing a benchmark for cross-dataset transfer learning in action recognition.

When utilizing appearance-based characteristics, the sign language recognition system must overcome a number of obstacles, including the visual variety of each sign language word's utterance, varying pronunciations, and a huge number of features de-

rived straight from the image frames. In isolated SLR, the use of transfer learning can be seen in cross-task transfer (from image recognition, action recognition [143], pose estimation [87] or continuous SLR [144]) or cross-lingual [27] settings. However, as noted in several studies such as [27] finetuning-based cross-dataset transfer between similarly sized isolated SLR datasets of different sign languages shows little to no benefit. In order to understand what information can be transferred in isolated SLR successfully with which method, a dataset that would allow us to experiment with closed and partial set transfer learning settings was necessary. Using such a dataset, it would become easier to independently test cross-dataset sign language transfer using several state-of-the-art transfer learning approaches. Such datasets exists for similar domains such as image classification [145], video classification [28], but not sign language recognition. As such a dataset existed for video classification with the HMDB [146] and UTF-101 [147] datasets, we follow similar procedures to construct a baseline subset of the BSign22k and AUTSL datasets.

3. SIGN LANGUAGE REPRESENTATION AND RECOGNITION

In this chapter, we analyse baseline sign language recognition (SLR) methods capable of recognizing isolated sign language glosses from a large vocabulary of signs. While there exist numerous studies on isolated SLR in the literature, recent state-of-the-art methods employed a fusion of deep learning techniques developed for the action recognition domain that use multiple modalities. These include 3D-CNNs with RGB and depth videos, skeleton coordinates with Graph-CNNs, and skeleton heatmaps with Temporal Representation Methods [16, 55, 87, 148]. As part of this study, we first introduce a dataset we collected to train and analyze isolated SLR models on TID. On this dataset, we implemented state of the art 3D-CNN based SLR methods. Next, we analyzed the shortcomings of such algorithms and detail our proposed ATAF architecture. Finally, we propose the fusion of 3D-CNN and ATAF method to obtain state-of-the art performance.

3.1. BosphorusSign22k Dataset

In this section, we present BSign22k, a new SLR benchmark dataset. [30]. The dataset was produced to assist both the linguistics and computer science fields. It includes (isolated) videos of Turkish Sign Language glosses from three distinct domains: Health, Finance, and common sign language words. This dataset contains videos produced by six native sign language signers, making it useful for user-independent sign language research. Figure 3.1 shows examples of signs from our dataset.

All of the sign video recordings in the dataset were obtained using Microsoft Kinect v2 [149] with 1080p (1920x1080 pixels) video resolution at 30 frames per second. Resolution of the frames is important for recognizing sign languages especially when interpreting the appearance information related to hand and finger regions. In addition, all of the videos have the same recording setup where signers performing the sign are

in front of a green background which is 1.5 meter far away from the camera.



Figure 3.1. Examples of signing performances from different signers in the BosphorusSign22k dataset.

Specifications of the BSign22k dataset are in Table 3.1. Since the dataset was collected with Microsoft Kinect v2 sensor, we provide RGB video, depth map and skeleton information of the signer for all sign videos in the dataset. Moreover, we also provide OpenPose [150] joints which include facial landmarks and hand joint positions in addition to body pose information.

The dataset has a vocabulary of 744 signs; 428 signs in Health, 163 signs in Finance domains, and 174 signs are commonly used words. In the scope of this thesis, two versions of the dataset were released namely BosphorusSign [29] and BSign22k [30]. In the second version we have improved several aspects of the BSign22k dataset; (1) reduced the number of unique sign classes by merging same signs, (2) merged all subsets, and (3) removed some of the signs and videos that had erroneous recordings. This has resulted in 744 sign videos, the length of which vary between 1 and 8 seconds.

3.1.1. BosphorusSign22k Phonological Labels

BosphorusSign22k is an isolated sign language recognition dataset with 744 labels. It contains signs representing a large variety of gestures. For this reason, when analysing sign language recognition results, changes in accuracy do not always correlate to an overall increase or decrease in performance.

Table 3.1. Specifications of the BSign22k Dataset.

Property	Description
Number of unique sign classes	744
Number of signers	6
Number of videos	22,542
Total Duration	∼19 hours (∼2M frames)
RGB Resolution	1920 x 1080 pixels
Depth Resolution	512 x 424 pixels
Frame Rate	30 frames/second
Body Pose Information (Kinect v2)	25 x 3D Keypoints
Body Pose Information (OpenPose)	25 x 2D Keypoints
Facial Landmarks (OpenPose)	70 x 2D Keypoints
2 x Hand Pose Information (OpenPose)	21 x 2D Keypoints

In order to allow for a detailed analysis of results based on each signs unique phonological properties, we identified and manually annotated four groups of properties as follows:

(i) Presence of the non-dominant hand: Each sign is classified as one-handed if the dominant hand is the only active hand during training. If the non-dominant hand also leaves rest pose and participates in the sign, the sign is classified as two-handed.

(ii) Presence of circular motion: The sign is classified as a circular sign if there exists a circular motion of hands or fingers for one or both hands during the performance.

(iii) Presence of repetitive motion: The sign is classified as a repetitive sign if there exists a repetitive motion of hands or fingers for one or both hands during the performance.

(iv) Number of morphemes: The sign is considered as monomorphemic if it only consists of a single handshape [151]. If the sign consists of a sequence of two or

more handshapes, it is classified as a compound.

Differences between the annotated properties of signs are illustrated in Figure 3.2. In the figure, the movement of repetitive, transparent and compound signs are given as overlapped transparent images to visually represent the differentiating types of motion in the sign.

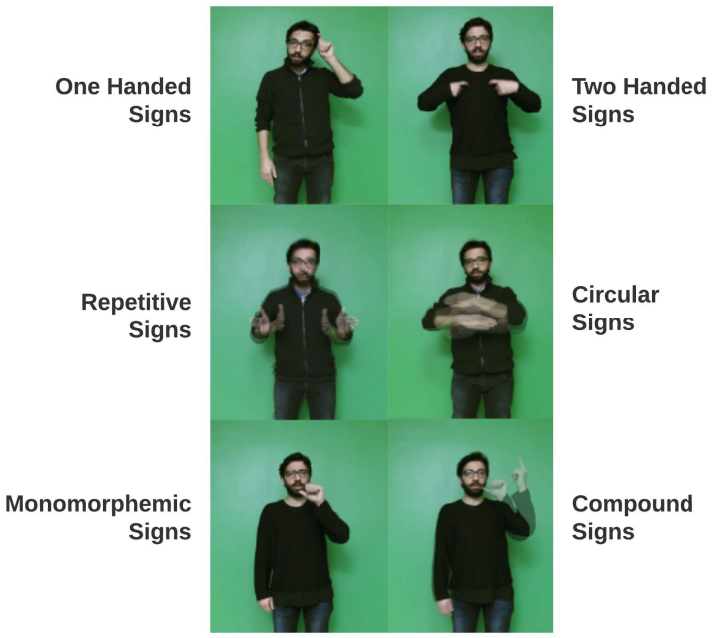

Figure 3.2. An overview of the analyzed sign types.

3.2. 3D Convolutional Neural Network Based SLR

In this section, we have used a video recognition method which is based on 3D ResNet architecture with mixed 2D-3D convolutions, also called MC3 in [74]. The model consist two residual blocks with 3D convolutions, three residual blocks with 2D convolutions and a fully connected layer as its classification layer. This network was built based on the hypothesis that learning temporal dynamics is beneficial in early layers while the higher levels semantic knowledge can be learnt in late layers [74].

To recognize sign language videos, we have performed experiments on finetuning different residual blocks of this model, which was pre-trained on the Kinetics-400 dataset. The proposed training method for this model used randomized clips of frames as its inputs, which is not suitable for our problem because randomized clips may include different non-recurring parts with the same isolated sign label, which will affect the training of the network. Thus, at the training phase, we randomly sampled batches of uniformly sampled frames from sign videos of the training set. This allows us to cover most of the sign video which is more suitable for our problem. While this approach can be extended with more advanced frame sampling methods such as keyframe sampling where we can obtain meaningful frames representing the sign, we used uniform sampling for our baseline results.

To evaluate baseline methods on the BSign22k dataset, we have used the publicly available implementation in PyTorch-torchvision [152] library for 3D ResNet model with mixed convolutions. For the training of the 3D ResNet architecture, we only focused on training the network from scratch and finetuning different layers of the pretrained networks on BSign22k dataset. During training, at each step, the network takes uniformly sampled 16 frame sign video clips with size of 112x112 pixels. Before feeding into the network, sign clips are horizontally flipped with probability of 0.5, and their pixel values normalized between 0 and 1. The proposed training approach [74] uses spatial random cropping of frames, but we discarded this step in our model training stage. After preprocessing, we train our network using Adam optimizer [153] with batch size of 32 sign video clips on a NVIDIA Tesla V100 GPU. For the testing step, we performed the same preprocessing approach as in training except horizontal flipping of frame clips in order to make a prediction from our trained model.Initial experiments on the BSign22k dataset with several different finetuning strategies are presented in Table 3.2.

Following the establishing of a baseline methodology through hyperparameter optimization, with freezing the first 2 Resnet block layers during finetuning, we turned our attention to selecting different modalities for better sign language representation.

Table 3.2. Baseline MC3 ResNet results on the BSign22k dataset.

Method	Top-1 Acc (%)	Top-5 Acc (%)
Training from scratch	57.76	84.22
Training only fc	55.03	81.98
Finetuning last 2 blocks	75.38	94.16
Finetuning last 3 blocks	**78.85**	**94.76**
Finetuning last 4 blocks	63.88	88.66
Finetuning all layers	71.02	92.51

Our assumption for this experiment was that 112*112 full-body images were not large enough to recognize finger configurations. Instead of increasing image size to 1024*1024 which would make our training slow, we opted to train 3 different classifiers focusing on hands face and full-body to obtain detailed representation for manual and non-manual features. Following the insight that the different cue models can capture a different subset of features, which can lead to better results when combined effectively, we train three models and apply fusion to obtain combined overall results. Standard fusion is applied by averaging softmax outputs as in [69]. In the weighted setting, we have applied weights to each model proportional to their validation accuracy via standard multiplication. Table 3.3 provides the result of the fusion.

The results of the fusion experiments are evaluated for three different settings where spatial and temporal randomization is applied during sampling of frames and crops. The weighted fusion of models belonging to three different modalities reach 94.94 % which achieves a 19.56 % accuracy improvement over the best results in Table 3.2. To analyse which cue benefits the fusion results the most, we have performed score fusion to all combination pairs of cue settings. According to this ablation study, we were able to observe the effect of each cue to the overall fusion. For example, to find the effect of the face model, we subtract the Body+Hand setting from the Body+Hand+Face setting. Table 3.4 shows the results of the ablation study.

Table 3.3. Classification accuracy results of the sampling and fusion settings. Three different settings are provided in the table. From left to right, (1) Single cue spatial sampling results, (2) Active Window Based Temporal Sampling applied to each crop, and (3) Spatial&Temporal settings are combined in one setting. Note that the bottom two rows include the fusion result of the above three models in each setting.

	Spatial		Temporal		S&T Combined	
Setting	Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5
Body	75.73	93.88	81.83	96.02	86.91	98.17
Hand	86.25	97.61	88.70	97.59	91.73	98.72
Face	24.27	44.45	37.00	57.89	39.12	59.33
Fusion	90.63	98.92	93.88	**99.65**	94.47	**99.78**
Weighted Fusion	**92.18**	**99.27**	**94.03**	99.56	**94.94**	99.76

Table 3.4. Effects of excluding individual cue units from the final fusion model. Using the different two cue settings and their performance, we infer to the excluded setting and its effect on the final mix.

Setting	Accuracy	Excluded Cue	Effect (%)
Body + Hand	91.80	Face	2.08
Body + Face	**84.66**	Hand	**9.22**
Hands + Face	88.70	Body	5.18

In our analysis, by performing an ablation study on the score fusion of all three approaches with accuracy 93.8%. We can see that hand cue has the most effect on the fusion by 9.22% which is followed by the body model with 5.18%.

To further analyze the types of signs in which the proposed method performs well and fails, we have labeled the 744 sign glosses in the dataset according to specific sign attributes. The sign classes are grouped into categories such as one-handed signs, two-handed signs, mono-morphemic signs, compound signs, and signs involving repetitive and circular movements of the hands.

Table 3.5. Analysis of the temporal sampling based recognition approach with respect to signs with certain grammatical sign attributes: one-handed signs, two-handed signs, mono-morphemic signs, compound signs, and signs involving repetitive and circular movements, respectively.

	Number of Classes with Selected Attribute								
	234	510	75	669	457	287	375	369	744
	One Hand	Two Hand	Circ.	Not Circ.	Rep.	Not Rep.	Mono	Comp.	All
Body	72.94	86.10	86.43	81.33	80.77	83.55	77.40	86.48	81.83
Hand	83.78	91.07	91.40	88.41	87.54	90.59	85.36	92.24	88.70
Face	45.33	33.01	29.41	37.82	36.39	37.99	33.79	40.52	37.00
Fusion	**91.82**	94.86	**96.15**	93.63	**93.45**	94.57	92.08	95.78	93.88
W.Fusion	90.87	**95.55**	95.70	**93.85**	93.37	**95.09**	**92.21**	**95.96**	**94.03**

Table 3.5 summarizes the analysis: The experiments are performed using temporal sampling with the best performing mixed convolution approach. Attribute-wise accuracy scores are calculated using the test set samples belonging to the classes containing the selected attributes. Overall, the accuracy scores in Table 3.5 demonstrate that for nearly all the subsets in the dataset, hand, body, and face-based features show consistency in their relative performance.

Looking at the results for different attributes one by one, we can see that signs involving two moving hands are better recognized than the one-handed sign glosses

in the dataset. The performance difference can be explained by the fact that in one-handed signs, the weight of handshape may be more critical than the two-handed signs. The relative positioning and appearance of both hands, which is more apparent, may be easier to represent for the neural network.

Secondly, compound signs have a greater recognition accuracy than monomorphemic signs (95.96% vs 92.21%). Considering the number of signing hands, the amount of additional information in the form of consecutive morphemes present in an isolated sign makes recognition easier, thus improving the performance system. From this result, we can infer that the method's representation power is higher when a sign is greater in length and contains different hand shape and position combinations.

Looking at repetitive gestures, we see a 1.6% improvement in accuracy when the signs do not contain repetitive hand gestures. The issue with repetitions, which we can attribute to this difference, is that the temporal and spatial forms of repetitions are more prone to differ between performances and users, in comparison to the static hand shape parts of the signs that follow specific rules.

Finally, we take a look at circular signs, which include circular hand and arm movements, which involve at least one entire rotation. These signs are dynamic signs where the hands do not stop while presenting a handshape. As these signs do not conform to the movement-hold phonological model of sign languages [36], representing them by choosing temporal frames is more complicated, reducing the effectiveness of keyframe based approaches [31]. Overall, the method performs well with circular signs, making fusion attempts with methods focusing more on the handshape of signs promising future leads.

4. LINGUISTIC REPRESENTATIONS FOR SLR

4.1. Aligned Temporal Accumulative Features

In Chapter 3, we used RGB based spatio-temporal architectures to model sign language signs in image sequences. While these methods achieved top performance on the BSign22k dataset these methods have do have some weaknesses. Ideally designed for the action recognition domain where context independent of a user is just as important, the information in sign language recognition lies solely in the users configuration of hands body and face.

Commonly used deep learning-based video recognition models often capture spatio-temporal representations by sub-sampling frames over time [74]. Large chunks of frames in these videos are repetitive and can be discarded to learn more concise features. In this work, we make use of two sign language features to find concise representations related to human upper body joint information; (1) we propose ATAF descriptors that are capable of accumulating joint coordinates over time to create fixed-sized compact representations from variable-length sign language frame sequences, and (2) we utilize multiple configurations of moving and static joints to infer grammatical keyframe information.

Like speech and written language recognition tasks, where boundaries between phonemes and words are clearly distinguished from each other, sign languages model transitions between phonemes and signs using grammatical rules. Using these rulesets, signers visually express themselves with a corpus of thousands of sign glosses, from simple commands to complex and abstract sentences. One of the major language models that handle transitions among phonemes and signs is the Liddell & Johnson [36] model where signs are formed as consecutive combinations of movement and hold subunits. While this is considered as a simplistic approach that generalizes over the intricacies of different sign languages belonging to different cultures, modeling the temporal as-

pect of sign language representations based on linguistic patterns is a popular research approach that has been successfully adopted by different studies [154, 155].

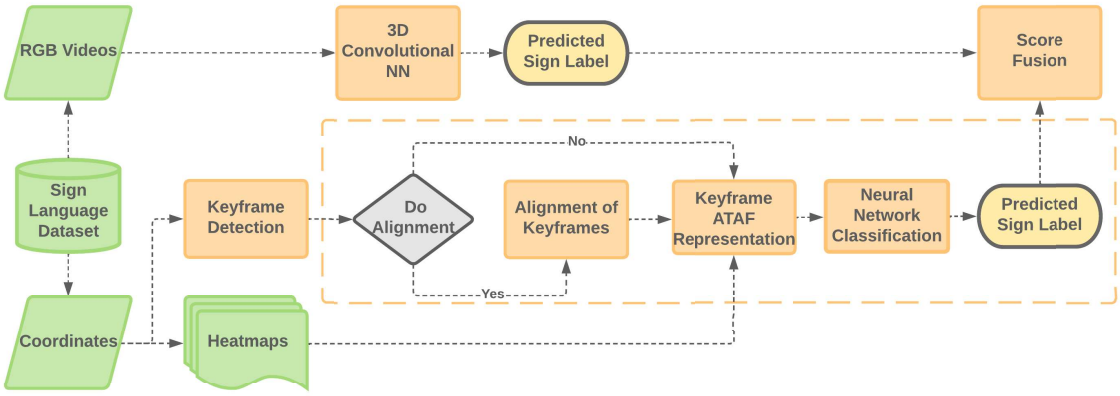

Figure 4.1. Overview of the aligned temporal accumulative features method. The proposed ATAF approach uses keyframe detection and keyframe alignment followed by sign language representation and classification. Predictions are then fused with the predictions of 3D-CNNs to improve recognition accuracy. Variants of ATAF without keyframe extraction and alignment are used to validate our proposed approach.

ATAF can be used to represent sign language videos as consecutive sequences of movements and holds, where keyframes are represented as static sign subunits. Through keyframe detection, we incorporate dynamic subunits, which are represented with varying colors based on their order of occurrence. Following keyframe detection and alignment, the final representations of gestures are composed of static keyframe information, temporal ordering of dynamic subunit frames, and joint heatmap strengths. 3D CNN based isolated SLR does not take into account *movements* where trajectories contain the linguistic information and *holds* where hand shape and location becomes critical in SLR. For this reason, we propose a spatio-temporal representation method that is robust to minor variations in SLR performance such as repetitions and signing speed changes. When fused with 3D-CNN based representations, the presented method yields results comparable to the state-of-the-art in isolated SLR task.

4.1.1. Method

Sign languages convey meaning through sequences of body part configurations. Our approach is inspired by the Liddell & Johnson [36] model, which treats signs as compositions of consecutive movement and hold parts. For modeling movement and hold patterns of sign languages (SLs), we introduce Aligned Temporal Accumulative Features (ATAF). In this section, we describe the components of ATAF, as illustrated in Figure 4.1: keyframe extraction, alignment, and accumulative feature extraction, followed by classification.

4.1.1.1. Keyframe Extraction. In this study, we use a keyframe detection approach which is based on hand joint speeds and density clustering, proposed by Tang et al. [106]. The approach has three steps; (1) calculating hand speed difference for the dominant hand, (2) performing density clustering to find local minima, and (3) detecting keyframes.

As preprocessing, frames from non-gesture segments are trimmed from the beginning and end of each isolated sign video. These frames are selected where both hands are below the hips and are stationary. Given a frame sequence belonging to a trimmed sign language video of length N frames, we extract a variable number of M keyframe candidates. In our experiments, our models require a fixed number of K keyframes. For all videos in the dataset, we choose a K value and extract K most suitable keyframes from the M keyframe candidates detected for a given video.

As we only want to extract keyframes from the dominant hand, we first perform dominant hand detection using joint displacement as a feature. J denotes the total number of joints for the dominant hand we use for keyframe detection. $d_{n,j}$ denotes the displacement of joint j between frames $n-1$ and n. $D_j = \sum_{n=1}^{N} d_{n,j}$ denotes the sum of all movement of a joint for a given gesture. Dominant hand selection is made by comparing the sum of wrist and fingertip joint speeds from all frames $\sum_{j=1}^{J} D_j$ for both hands. The hand with a larger sum of joint displacement over the entire video is

selected as the dominant hand.

We then use the dominant hand information to infer candidate keyframes. From each video, we detect a variable number of M candidate keyframes, which are local minima values for displacement $d_{n,j}$. We select a list of candidate keyframes and save their speeds $P_{m,j}$ as

$$P_{m,j} = \{d_{n,j}, \text{ if } d_{n,j} < d_{n+1,j} \text{ and } d_{n,j} < d_{n-1,j}\} \tag{4.1}$$

where proximity thresholding is used to filter out consecutive detections from different joints. In this filtering, the frame n with the greater total joint movement $\sum_{j=1}^{J} D_{n,j}$ is discarded. The remaining extrema points are then ranked according to their relative surroundings. The points with the highest density difference from their surroundings are selected as the strongest keypoint candidates.

Density clustering proposed in [107] is an optimal method for detecting these points. For each candidate point s, local density δ_s and the distance ϕ_s from the closest point k with a higher density are calculated for each joint as proposed by Tang et al. [106] as in Equation 4.2. f_{P_s,P_k} is the point distance between two points while f_c is the cutoff threshold. $Z(z) = 1$ if $z < 0$ and $Z(z) = 0$ if $z >= 0$. This function yields the number of points closer than the distance f_c to the point P_s. Through inspection, the cutoff threshold was determined as a small number to increase the number of detected candidate keypoints. The candidate keypoints are calculated using

$$\delta s, j = \sum_{P_k} Z(f_{P_s P_k} - f_c) \tag{4.2}$$

$$\phi_s = min_{\delta_k > \delta_s}(f_{P_s P_k}) \tag{4.3}$$

where we choose the K keypoints from the suggested M points with the largest ϕ_s values. In the implementation, we use two sets of joints for calculating keyframes. The wrist speed-based keyframing method uses the hand speed of the dominant hand to detect keyframes, while the finger speed method makes use of the fingertip speeds of the thumb, index, and middle fingers as well as the hand base speed of both hands to detect more keyframes. As the second approach detects more keyframes than the first method, a pruning of the keyframes is done using the ϕ_s values and distances of the

detected keyframes. In the pruning method, first keyframe detections that are at least 3 frames apart are eliminated by choosing the strongest detection. Of the remaining ones, K keypoints with the largest ϕ_s values are chosen.

4.1.1.2. Alignment with Temporal Transformer Networks. Due to personal gesturing style or differences in linguistic syntax, there may be variabilities in the occurrence of keyframes of a sign. As seen in Figure 4.2, these irregularities leave us with a different number of keyframes, which may be brought into correspondence by alignment. However, while bringing similar frames of the same signs into correspondence, alignment must not prevent the differentiation of distinct signs. In the ATAF representation, this is achieved by incorporating different alignment approaches with the TTN method presented in [120].

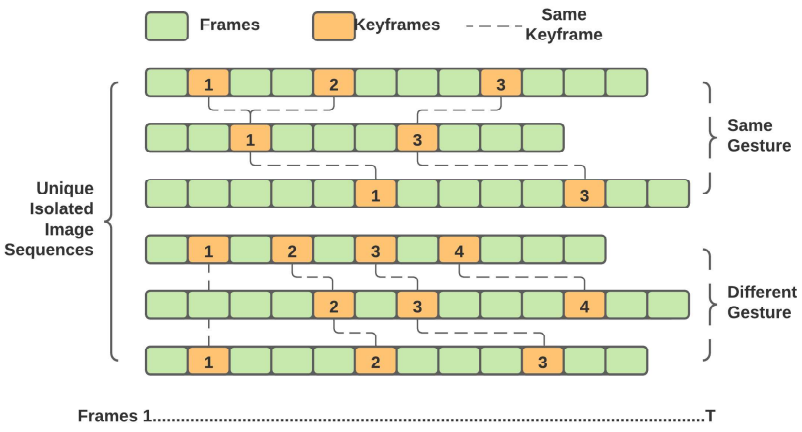

Figure 4.2. A visualization of keyframe based alignment. Given sequences of different length belonging to different classes, candidate keyframes from each sign are aligned to create better temporal representations.

TTNs provide a framework for end-to-end temporal alignment and classification with neural networks. TTN module, which is differentiable, can be used inserted in front of a classification network. For an image sequence of length T, representing a

sign language video as $I = \{i_1, i_2, \ldots, i_T\}$, our proposed alignment approach makes use of numerical features representing the gesture in every frame such as coordinates $X_{D,T} = \{x_{D,1}, x_{D,1}, \ldots, x_{D,T}\}$. The method finds a warping function γ, which can be used to obtain the alignment $\eta(I, \gamma)$, $\eta(X_{D,T}, \gamma)$ or any other temporal feature obtained from I such as ATAF. The module is capable of reducing intra-class variance by generating input-dependent warping functions, which lead to representations that have improved capabilities in representing compound and repetitive gestures due to improvements in the modeling of consecutive morpheme sequences.

The alignment approach consists of three groups of layers. These are the trainable layers, constraint satisfaction layers, and the warping layer. For the training layers, we have implemented two different approaches that utilize different types of inputs. The first set of training layers take $X_{D,T}$ the coordinates for several hand joints as input. This input is subjected to 1D convolution layers, batch normalization, 2D adaptive max pooling, and fully connected layers to obtain a feature sequence of length T. The second approach for the training layers uses the same heatmaps we use for ATAF features as inputs. *Nose*, *Hand*, and *Finger* heatmaps are used as the inputs. The method first performs temporal 1D convolutions followed by the application of temporal attention. The temporal attention vector is calculated using normalized hand speed for the dominant hand for each frame. At this stage, the features are shuffled, and 2D convolutions with a kernel size of 3×3 are applied. Finally, adaptive average pooling is applied to obtain a 1D feature vector of length T. The calculation of both methods is presented in Figure 4.3.

The constraint satisfaction layers take as input an unconstrained vector and converts it to a warping vector that satisfies boundary conditions as well as the condition for the function to be monotonically increasing such that $\gamma(0) = 0$, $\gamma(1) = 1$ and $\gamma(t_1) < t_2$ if $t_1 < t_2$. To this end, we first apply L2 normalization as

$$\gamma = \frac{v}{\|v\|} * \frac{v}{\|v\|} \tag{4.4}$$

to the vector v. As described in [120], we use the element-wise product of the resulting matrix with itself and obtain cumulative sum values for each time frame. The result

is multiplied by video length T and rounded to the nearest integers as

$$\gamma(t) = T \cdot \sum_{i=1}^{t} \gamma(i) \tag{4.5}$$

to obtain the final warping function $\gamma(t)$ at time step t.

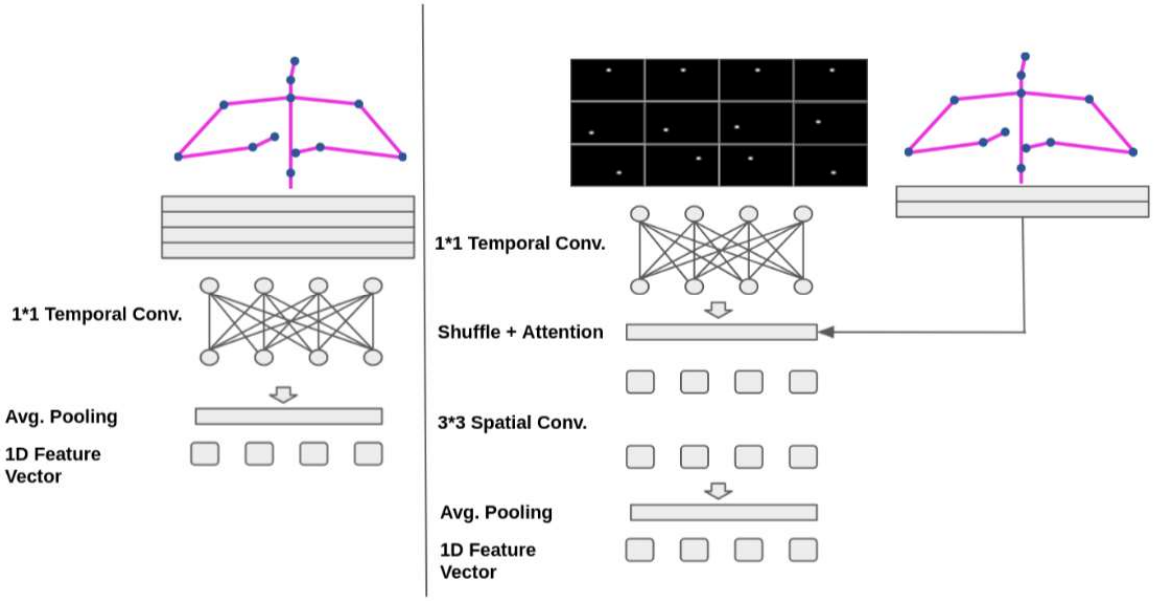

Figure 4.3. Different keyframe feature representations to calculate alignment. On the left side, hand coordinate based features are processed, while on the right side heatmap-based features are presented.

Finally, in the warping layer, the input sequence is warped using the function γ. To warp a sequence of length T, an index function is calculated as $\Gamma(T) = round(\gamma(1), \gamma(2), \ldots, \gamma(T))$. The indices calculated by the ordering may not correspond to integers. Rounding is performed following the linear interpolation to length T. This index function can then be used to reorder the frames of $I = \{i_1, i_2, \ldots, i_T\}$ and any temporal features derived from these frames. These reordered frames/features are then trained with a classification network. Since both networks are trained with classification cross-entropy loss, the learned alignments aim to maximize classification accuracy.

In terms of choosing what to align, we use the proposed alignment approach to align either all the frames of a given gesture or linearly sampled frames or keyframes. For both approaches, during network training, two different alignment approaches can be employed. In the standalone alignment case, the output of the warping function γ is given to a neural network to perform classification with cross-entropy loss. After the optimal warping is obtained, ATAF features are computed with the learned warping to perform SLR. In the second case, end-to-end ATAF classification is performed by combining TTN alignment, ATAF feature re-colorization, and neural network classification stages.

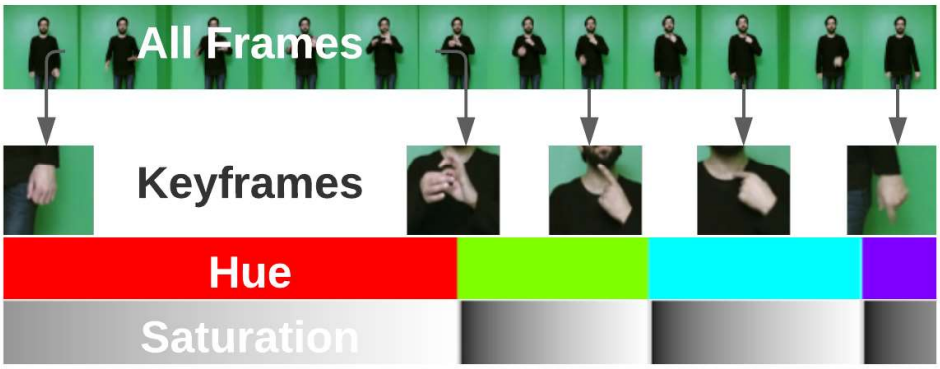

Figure 4.4. Overview of the HSV based coloring method using keyframes. Each keyframe is uniquely color coded with a unique hue channel and a gradient style saturation channel indicating temporal location. The final feature is calculated by stacking joint heatmaps, applying aggregations and converting the resulting image to RGB format.

4.1.1.3. Temporal Accumulative Features. Using 2D joint coordinates as inputs, these features can characterize the motion and trajectory shape of individual or combinations of joints such as fingers, hands, arms, and body. These joints are represented as heatmaps and converted to accumulated feature images to be fed into 2D CNNs.

ATAF approach utilizes keyframes and skeleton joints to represent subunits of SL gestures. Since SLs are multi-channel languages, using the locations of human body parts such as hand and finger joint coordinates is useful for describing the motion of signers. Our approach accumulates these joint features over time to create fixed-size compact representations from variable-length SL gesture videos.

Heatmap images are converted to color images as in Figure 4.4. The hue channel represents different keyframes where a unique color is assigned to each keyframe. The saturation channel represents time spent until a new keyframe is encountered. This way, the ATAF approach can be used to represent SL videos as a consecutive sequence of movements and holds, where keyframes approximate static sign subunits. The visual representation of an isolated gesture with the accumulated features is shown in Figure 4.5 for the right-hand joints.



Figure 4.5. The ATAF colorization for the left hand joints are shown in the left image. In order to represent all joints, similar feature images belonging to all critical joints are given as inputs to the learning model.

For each joint, an accumulated heatmap image is produced by convolving the joint locations with a gaussian kernel, forms the value channel. The hue and saturation channels of ATAF are created (illustrated in Figure 4.4) using temporal static subunits, which may be linearly sampled frames or keyframes detected with the length of T. The $T \times h \times w$ first channel H_j contains temporal subunit identifiers for a sequence of T

frames for joint j. In contrast, the second channel S_j contains temporal ordering indices of frames in the range [0,1] for each subunit.

Having obtained the accumulated heatmap image V_j, we normalize it by dividing it with the maximum value for each joint $\max(V_j)$ as

$$V_j^{'} = V_j / \max(V_j). \tag{4.6}$$

Next we construct the temporal segment indicator vector H and the temporal frame indices vector S. Following keyframe or temporal segment extraction, we obtain the beginning and end temporal locations for each video and use it to construct these vectors. Given an image with K keyframes $K = k_0, k_1, ...k_{K-1}$, H is a monotonically increasing T length vector whose elements are in the range $(0, K-1)$. Each value of the vector H denotes which keyframe the joint belongs to at a given temporal timestep. S is a T length vector whose elements are in the range $(0, 1)$. The vector element values are set to 0 at the first frame of each keyframe and linearly increase to 1 which is assigned to the corresponding keyframes last frame. Example vectors for H and S can be constructed as: $H = \{1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2 ....(K-1), (K-1), (K-1)\}$ and $S = \{0, 0.25, 0.50, 0.75, 1, 0, 0.25...., 0.75, 1\}$.

Following the construction of H and S, these vectors are replicated into matrices for each frame and joint and then element-wise multiplied by $\lceil V_j \rceil$ and averaged over T frames. The multiplication with the sparse matrix V_j makes the H_j and S_j matrices sparse as well. The resulting images are combined as

$$I_j = H_j + S_j + V_j^{'} \tag{4.7}$$

$$U_j = [H_j; S_j; V_j^{'}] \tag{4.8}$$

$$N_j = \frac{U_j}{\mathcal{E} + I_j} \qquad j = 0, 1, 2, \ldots, 29 \tag{4.9}$$

where dividing each line's maximum with the sum of all pixels over time yields a height by width image. In this image a pixel being brighter for a joint means the joint stays in the same pixel for a greater duration.

In order to normalize this feature with respect to the amount of time spent in a pixel, I_j is used to obtain N_j in Equation 4.9, which gives the trajectory of a joint in a

sequence-independent of the joints speed, thus focusing more on the traveled path than execution speed for the first feature U_j. $\mathcal{E}$ denotes the smallest float value to avoid division by zero. The concatenation of U, N and I yields a $7 \times h \times w$ sized feature for each joint, which yields the combined ATAF feature. In the final step of feature calculation, U and N matrices are converted from HSV to RGB to linearly combine temporal keyframe information with joint heatmaps. The features for a group of hand joints are visualized in Figure 4.6.

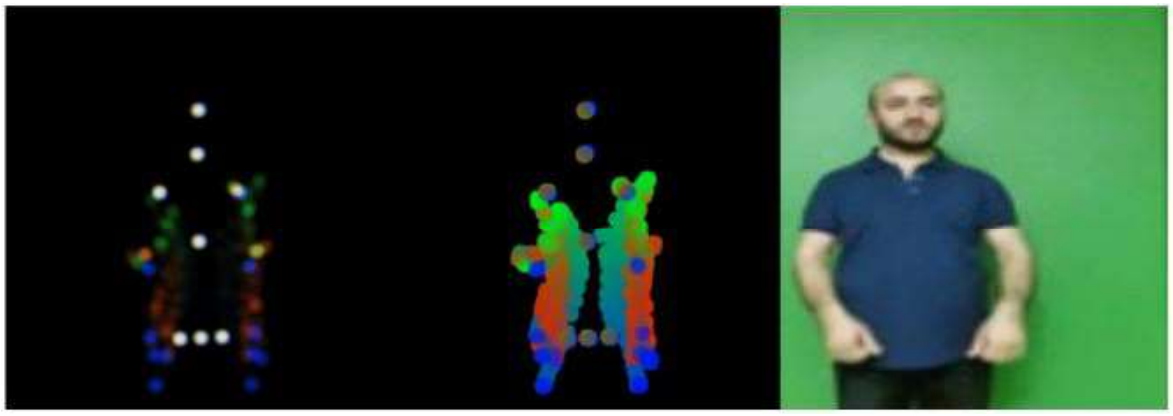

Figure 4.6. The features U_j (left) and N_j (center) are visualized for a group of hand and body gestures.

For the task of SLR, in addition to calculating joint heatmaps for every joint, combined joint heatmaps are calculated from body parts such as hands and upper bodies. Since SL gestures tend to show large intra-class variances as well as inter-class similarities, the relative locations of trajectories in the same action groups contain differentiating information. Features for five different action groups are used for these groupings as *Head* (*Head*, *Neck* and *Torso*), *Arms* (*Shoulders*, *Elbows*), *Wrists*, *Hands*, *Finger Bases* and *Finger Tips*.

4.1.1.4. Classification. Following ATAF feature calculation, we represent each isolated sign language video with ATAF representations with the size of $H \times W \times J \times 7$,

where $H \times W$ is the size of the heatmap and J is the number of joints used in the construction of the feature. The 7 channels in these features are U (RGB Image), N (RGB Image) and I (Intensity Image) matrices explained in section 4.1.1.3. The ATAF features are classified with a shallow and fast network with random initialization as the heatmap-based features do not benefit from prior learned weights. The network consists of a mix of convolution, pooling, and fully connected layers.

The model has two blocks of layers in the recognition architecture, each with two convolutional layers of strides 2 and 1, respectively. After each block, we doubled the number of filters of convolutional layers to reduce the spatial dimensions. We have also applied batch normalization and non-linearity after each convolution layer. We apply global average pooling before the fully connected layer for reducing the number of parameters of our model to prevent overfitting. As a result of training hyper-parameter optimization experiments, a shallow architecture with three blocks and two convolution layers per block was found as the optimal architecture for ATAF. Details on the architecture selection process for heatmap based features can be found in [14, 156]. Training optimizations also showed that as the number of parameters increased further, the models started to overfit and not converge. We stack Temporal Accumulative Features over selected six joint groups from body and hand joints and fed them into the network as inputs. The model is trained with a cross entropy loss function and classification predictions are obtained for class predictions.

Combining the representation and classification power of different classifiers with different strengths and weaknesses has been a popular method for building meta classifiers with increased accuracy. In order to combine the proposed model with models trained using different modalities such as 3D-Convolutional models [74], we use score level fusion as suggested for neural network based architectures by [69]. Since both classifiers are tested on the same videos with the same class outpus, We utilize score level fusion via averaging neural network outputs from both networks. We concatenate the logit outputs of each sample for each classifier and apply elementwise mean operation along the same concatenated dimension. Softmax function is applied to the

resulting outputs for each sample to obtain top-k predictions of the classifier.

4.1.2. Experiments

In this section, we present a detailed ablation study on representation options, choice of skeleton joints, keyframe selection, sampling, and alignment. We carry out the ablation experiments on subsets of the BSign22k dataset. After finalizing the methods, we present comparisons with state-of-the-art skeleton-based techniques (SL-GCN [16], HCN [55] and SSTCN [16]) on two datasets: BSign22k and AUTSL.

Our experiments were conducted on the BSign22k [10] and AUTSL [58] datasets.In the experiments of this study, we use the training and validation splits from [10], where the examples from 5 signers are used for training and samples from one signer is used for testing, thus performing user independent experiments. The dataset has various subsets that contain 25 (25C), 100 (100C), 174 (174C), 744 (744C) classes. The 174 class subset differs from the other subsets as its contents are selected according to the semantic gesture category, i.e., general-purpose words of the BSign22k dataset. The rest of the dataset splits' contains signs from general-purpose as well as health and finance domains as reported in [10]. To speed up the convergence of numerous ablation experiments, we perform parameter optimizations with smaller data subsets and extend to larger subsets to make final evaluations on the best-performing models. In addition, the dataset has labels for various mechanical sign properties such as one-handed, two-handed, repetitive, and circular that can be used to evaluate the developed methods.

In order to perform a detailed performance analysis of the gestures with respect to sign language grammatical attributes, we grouped certain signs of the BSign22k dataset. Of these groups, One-Handed vs. Two-Handed signs separates the dataset into two clusters based on the number of hands used in a sign. Likewise, a sign is repetitive if the gesture contains repetitive movement and circular if the sign involves circular movement. In terms of sign composition, mono-morphemic signs are composed

of a single morpheme (smallest meaningful unit), while compound or multi-morphemic signs are made of multiple morphemes.

To verify our results on an additional dataset, we performed experiments on the AUTSL dataset [58]. The dataset contains isolated sign language videos of 226 unique signs belonging to Turkish Sign Language. It contains 38,336 videos belonging to 43 different signers. The videos all contain a single person in a standing or sitting position performing gestures. As a result, accurate Openpose finger coordinates are obtained for all videos allowing SLR with the methods proposed in this paper. The results on the dataset are obtained by training the dataset on the provided training split, and the test results are reported for the validation split of the dataset.

For all experiments, we implement our experiments in Pytorch [152] and use KinectV2 and Openpose libraries for pose estimation. Training and evaluations of the method are performed on a cluster with V100 GPUs. We train the ATAF models with cross-entropy loss as well as AdamW [157] optimizer with an initial learning rate of 0.0001 and weight decay of 0.0005. The training was set to 200 epochs for smaller subsets and 800 epochs for the 744 class models, respectively, with a batch size of 8 and a crop size of 320×180. For data augmentation, we use random scaling and cropping as well as temporal random sampling of frames. Experiments showed that data augmentation techniques such as horizontal mirroring and randomly translating coordinates improve recognition performance. Other techniques such as random shifts in duration or coordinate normalization did not result in beneficiary effects.

4.1.2.1. Analysis of Accumulative Representation Methods. In Table 4.1 the results of comparative analysis of recognition between the proposed approach and comparable Potion method is presented.

Table 4.1. Effects of the number of linearly sampled temporal subunits parameter on recognition accuracy (%).

Aggregation Method	Subunit	174C	100C
Potion [14]	2	72.2	71.6
	3	73.3	72.6
	4	70.7	71.6
	6	64.5	68.8
	8	57.2	51.6
ATAF	2	72.6	51.6
	3	**75.1**	78.8
	4	72.8	**79.1**
	6	73.6	72.6
	8	67.5	52.4

In both approaches, ten Kinect joints belonging to the *Head*, *Arms*, and *Hands* construct heatmaps. These heatmaps are temporally accumulated to construct sign representations. Both methods use linear temporal sampling to select frames. The number of linearly sampled temporal subunits parameter determines the amount of temporal chunking in the representation by modifying the total number of color channels in Potion and the number of hue-groups in ATAF. For the sake of comparability, the results in Table 4.1 exclude the use of keyframing and alignment improvements for ATAF. Overall, the ATAF method seems to yield better results than the Potion approach in both 174 and 100 class datasets.

4.1.2.2. Ablation Study on Feature Representation Schemes. This subsection analyzes the components that make up the Temporal Accumulative Features and finds the optimum representation. The ATAF feature is composed of three major components, the joint heatmap and its aggregation (value channel), the temporal segment

indicator (hue channel), and the temporal anchor (saturation channel). In Table 4.2, we analyze different combinations of image and Gaussian kernel sizes used to generate the heatmap images from joint coordinates.

Table 4.2. Recognition accuracy (%) of different image and kernel sizes.

Image Size	Joint Heatmap	50C
480x270	Point	81.62
	3x3 Gaussian	82.71
	5x5 Gaussian	84.56
	7x7 Gaussian	84.01
320x180	Point	86.15
	3x3 Gaussian	87.36
	5x5 Gaussian	88.05
	7x7 Gaussian	88.44
240x135	Point	86.79
	3x3 Gaussian	88.44
	5x5 Gaussian	**88.87**
	7x7 Gaussian	**88.87**

Due to the large memory requirements of the Potion method, the ablation study experiments are performed with the first 50 class subset of the BSign22k dataset using ATAF with an equal length of 5 samples. The Top-1 accuracy results demonstrate that resizing the coordinates and projecting them onto smaller images increased the feature accuracy, contrary to expectations. Likewise, using larger gaussian kernels up to 5×5 increased accuracy, while further increases were not beneficial in small images. As a result, we found a balance between speed and accuracy by using small images. The model architecture for ATAF used in these experiments was determined through a similar experimentation strategy with [156]. In our experiments, we optimze the problem and define parameter values for the BSign22k dataset. We apply the

same parameters for AUTSL as experiments with minor changes in the best choice of parameters of BSign22k did not cause large variations in results. In Table 4.3, we demonstrate several different colorization schemes in representing gestures.

Table 4.3. Ablation study of ATAF with different feature compositions. The displayed results are accuracy (%) values on the first 50 classes of the BSign22k Dataset.

	C=3	C=5	C=7
Proposed	86.37	**88.87**	**88.87**
HSV	65.16	63.12	67.15
No Hue	80.36	80.36	80.36
Max Hue	82.25	82.86	83.25
No Saturation	85.98	87.92	88.12
Max Saturation	86.37	88.45	88.62
Normalized Coordinates	**86.56**	88.43	88.11

The proposed approach results presented in the first row include outputs where the temporal segment indicators, temporal anchors, and accumulated heatmaps are combined to form an HSV image. The temporal segment indicators and temporal anchors are temporally accumulated by calculating their per-pixel average values for each isolated video. The outputs are then converted to two separate RGB images for each joint. To contrast the baseline method, the HSV row shows the original ATAF output's accuracy without the conversion to RGB of the finalized feature. Conversion to RGB performs significantly better as the linear combination of all three channels creates a more concise descriptor from which the model can learn. Secondly, we compare the average aggregation of temporal unit information in the hue channel with a feature that excludes it (No Hue) and with another feature that obtains the per pixel maximum value over time. The average value is better in all three cases. Next, we make the same comparison for the temporal anchor in the saturation channel with no saturation, maximum saturation, and average saturation values over time. The modifications fail to beat the temporal per-pixel averaging approach present in the proposed approach

in both cases. Finally, we experiment with normalizing human joint coordinates by aligning the mid-hip and mid-shoulder points of all skeletons. The normalizations do not provide a meaningful improvement over baseline results for the best cases where the temporal segment values are more significant than 5 and 7. Therefore, they are excluded from further experiments.

4.1.2.3. Hyper-parameter Optimization for ATAF. Next, we searched for the optimal neural network complexity to represent isolated sign language videos on the BS50 subset in Table 4.4.

Table 4.4. Neural network parameter optimization for Hue based temporal colorization without temporal subunits.

Dropout	Blocks	Convs	Filters	Accuracy	Top-5
0.25	1	2	128	88.2	96.5
0.25	2	2	128	96.1	98.2
0.25	3	2	128	96.7	100
0.25	4	2	128	92.3	97.1
0.25	3	1	128	85.8	96.2
0.25	3	2	128	96.7	100
0.25	3	3	128	95.4	98.3
0.25	3	4	128	91.2	97.6
0.25	3	2	128	96.7	100
0.5	3	2	128	94.6	96.2
0.75	3	2	128	75.6	91.9
0.25	3	2	64	94.5	98.5
0.25	3	2	128	96.7	100
0.25	3	2	256	96.6	100

The experiments demonstrate that a shallow architecture with three blocks and two convolution layers per block yielded the optimal experimental results. As the number of parameters increased further, the models started to overfit and not converge.

4.1.2.4. Choosing the Best Set of Joints for TAF. In Table 4.5, we evaluate different joint combinations among the combinations of joints obtained from Kinect and Openpose libraries. The experiments are performed on the first 50 classes of the BSign22k dataset.

Table 4.5. The 25-class accuracy of the linearly sampled ATAF for different joint groups is shown. For each joint group shown as JG# joints belonging to the action group are included in the feature. The S and G columns indicate single joints being added as individual features vs groups of joints in a single feature.

	Kinect Body Joints								Openpose Hand Joints						25C	
	Head		Arms		Wrists		Hands		Wrist		Base		Tips		Top-1	Top-5
	S	G	S	G	S	G	S	G	S	G	S	G	S	G		
JG0	X		X		X		X		X				X		96.15	100
JG1					X										45.51	91.02
JG2	X				X										51.28	94.87
JG3	X		X		X										58.97	92.94
JG4	X		X		X		X								83.97	100
JG5	X								X						63.46	99.35
JG6	X								X				X		95.51	100
JG7	X								X		X		X		95.51	100
JG8	X	X		X		X									50.64	96.79
JG9	X	X								X				X	61.53	96.79
JG10					X					X					74.35	99.35
JG11		X	X	X		X			X				X	X	96.15	100
JG12	X								X				X	X	83.3	93.35
JG13	X	X		X					X	X			X	X	94.23	98.71
JG14	X	X	X	X					X	X	X		X	X	97.43	100
JG15							X								68.58	99.35
JG16	X						X								74.99	98.71
JG17	X		X				X								82.69	100
JG18	X						X						X		93.58	98.71
JG19	X						X				X		X		96.15	99.35
JG20	X		X				X				X		X		96.15	100
JG21	X		X				X						X		96.15	100
JG22	X		X				X				X				92.94	100
JG23	X		X						X		X		X		96.15	100

In the experiments, joints are grouped into action units depending on the source and types of joints. The experiments show that a combination of head and arm joints from the Kinect sensor and the Openpose joints yield high accuracy. Besides, grouping several joints into one feature image was also beneficial in further improving overall accuracy. Such an increase can be observed by the increase in accuracy from Joint Group 23 to 14. The experiments showed the importance of using finger joints from Openpose as the top performance without such joints was reported as 84%.

4.1.2.5. Using Keyframes to Colorize Temporal Video Segments. In these experiments, we analyze the use of detected static keyframes to segment videos instead of linear equidistant or random intervals temporally. The keyframes are detected by using a density peak clustering approach from different hand joints.

Table 4.6. Top-1 accuracies (%) of approaches that use equal length temporal segments (TS) or keyframes detected with *Hand* and *Finger* speeds. The experiments are performed on the first 50, 100 and 744(all) classes of BSign22k dataset.

Approach	# of TS	50C	100C	744C
Random Sampling	3	82.6	-	-
	5	**83.4**	71.86	64.16
	7	81.6	-	-
	9	80.9	-	-
Linear Sampling	3	88.29	-	-
	5	**88.87**	77.12	73
	7	87.29	-	-
	9	87.54	-	-
Wrist Keyframes	3	93.96	-	-
	5	94.22	-	-
	7	94.26	-	-
	9	**95.16**	**81.36**	**77.52**
Finger Keyframes	3	93.40	-	-
	5	**95.41**	80.89	75.92
	7	94.85	-	-
	9	94.80	-	-

We experiment with two sets of joints for hand detection: the first one involves the wrist joint of the dominant hand and the second one additionally includes the tips of all five fingers. The first method uses the coordinates of the left and right arm wrist joints, while the second approach uses the coordinates of the tips of the thumb, index, middle fingers, and wrist joint to capture finger movement. The use of movement change from more joints to detect keyframes results in more keyframes being detected in the finger-based keyframe method compared to the one only using wrist joints. In Table 4.6, we compare the two keyframe detection methods with two approaches where consecutive temporal groupings are selected independent of hand movement speed and sign linguistics. In the linear sampling-based approach, equal length subunits are chosen, while in the random sampling approach, random length consecutive subunits are chosen and used for ATAF generation.

The accuracy figures in Table 4.6 show that both keyframing methods have comparable performance. We experiment with different number of temporal segments in the 50 class subset and use the best temporal segment on the larger subsets. The ideal number of keyframes seems to fluctuate between 3-7 keyframes. The main difference between the wrist and finger-based keyframes is that the finger-based keyframing method finds more static frame candidates as peak values from multiple fingers are concatenated. Some gestures may contain only finger movement while the wrist stays constant, causing more true positive detections and detecting more false positives as keyframes. However, the wrist-based keyframe approach is seen as the method with the highest 744 class accuracy. The difference in performance can be explained by the extra keyframe candidates causing larger differences among the same class or keyframe detections where the wrist stays constant and the fingers move having an adverse effect on overall recognition performance. In order to figure out this issue, we introduce keyframe alignment to the model in the next experiment

4.1.2.6. Keyframe Alignment for Better Temporal Modeling. In this experiment, we try to learn an alignment for the keyframes detected in the previous section. The keyframes are aligned in an end-to-end classification setting to maximize recognition

performance. In Table 4.7, the unaligned wrist and finger keyframes from Table 4.6 are aligned with TTN to increase representation effectiveness.

Table 4.7. Comparison of recognition accuracy (%) on keyframe based ATAF effectiveness with and without TTN. The experiments are performed on the first 50, 100 and 744(all) classes of BSign22k dataset.

Alignment	Keyframes	# of KF	50C	100C	744C
Linear Sampling	Wrist Keyframes	3	93.96	-	-
		5	94.22	-	-
		7	94.26	-	-
		9	**95.16**	81.36	77.52
Linear Sampling	Finger Keyframes	3	93.40	-	-
		5	**95.42**	80.89	75.92
		7	94.85	-	-
		9	94.73	-	-
TTN KF	Wrist Keyframes	3	94.52	-	-
		5	95.22	-	-
		7	94.52	-	-
		9	**96.31**	83.71	78.09
TTN KF	Finger Keyframes	3	92.70	-	-
		5	96.54	-	-
		7	95.84	-	-
		9	**96.58**	**83.96**	**80.14**

The results demonstrate that as the number of keyframes increases, so does the benefit of aligning keyframes. Since the duplicates increase in the finger joint-based keyframe detection scheme, the alignment of gestures is most beneficial. Maximum accuracies are obtained with the alignment of 5 and 9 finger keyframes per gesture.

4.1.2.7. Alignment of Linearly Sampled Frames with TTN. One drawback of keyframe-based alignment is that its accuracy is dependent on the success of keyframe detection. To overcome the said drawback, we implemented a second alignment-based approach

where T linearly sampled frames from each gesture are aligned with each other. The proposed approach allows direct comparison with the keyframe-based model when the number of selected frames is low. The results can be observed in Table 4.8.

Table 4.8. Analysis of the Temporal transformer based method without using keyframe selection. Linear sampling with different number of frames (# of frames) are used to select (ts) number of temporal segments.

Alignment	# of ts	# of frames	50C	100C	744C
TTN Linear Sampling	3	20	**94.21**	-	-
	3	40	94.18	-	-
	3	60	93.15	-	-
	5	20	95.19	-	-
	5	40	**95.21**	**84.01**	**78.22**
	5	60	95.11	-	-
	7	20	94.19	-	-
	7	40	**94.59**	-	-
	7	60	94.47	-	-
	9	20	**93.91**	-	-
	9	40	93.79	-	-
	9	60	**93.91**	-	-

Table 4.8 contains parameters for determining the number of aligned frames and temporal segments utilized to color the output vectors. In keyframe-based alignment, we aligned sequences of keyframes detected from gestures. As the majority of signs in our dataset did not result in the detection of more than ten keyframes, aligning more frames was not possible with that approach, as seen in 4.7. In this approach, we align linearly sampled frames to find a similar number of temporal segments. We choose the number of frames to align as 20, 40, and 60, respectively. The optimal numbers of frames to align are 20 and 40 while using five or temporal chunks to colorize uniquely also yields the best performance.

4.1.2.8. Fusion Performance in Terms of Sign Attributes. The previous experiments on Tables 4.8 and 4.7 showed the optimal parameters for keyframing concerning accuracy over the entire dataset. In order to further analyze the strengths and weaknesses of the sampling and keyframing-based alignment approaches, we take the models analyzed in the previous section and report individual performances on different gesture sets based on the gestures' syntactic properties: We carry out tests on the one-handed, two-handed, circular, non-circular, repetitive, non-repetitive, monomorphemic, and compound signs. In Table 4.9, we see that while keyframe and sampling-based alignment methods perform comparatively well for two-handed signs than one-handed gestures, either method does not significantly differ in terms of their ability to capture signs based on the number of hands involved in the sign.

Table 4.9. The accuracies (%) of Linear sampling, the keyframe-based method, TTN Alignment of Keyframes and TTN alignment of linearly sampled frames are reported on different gesture subgroups: One-handed (1H), two-handed (2H), circular (C), non-circular (NC), repetitive (R), non-repetitive (NR), monomorphemic (MM), and compound (COM) signs.

	234 1H	510 2H	75 C	669 NC	457 R	287 NR	375 MM	369 COM	744 all
Lin.Sampling	67.35	76.23	69.43	71.20	73.15	67.63	66.12	76.00	71.02
Keyframes	74.16	**82.15**	73.15	**76.82**	74.67	79.28	68.44	84.59	76.45
TTN Lin	75.25	77.36	**75.75**	76.55	**78.10**	73.62	**71.99**	80.82	76.37
TTN KF	**75.52**	78.49	72.15	77.68	75.12	**80.30**	69.15	**85.22**	**77.12**

The linearly sampled scheme performs significantly better with circular and repetitive signs. In contrast, the keyframe-based approach is better at capturing compound signs. The analysis of repetitive and circular signs was chosen in part due to their unique movement schemes. For the circular signs, slowdowns in motion, which can be picked up as keyframes, do not occur. In contrast, for repetitive signs, more than one slowdown occurs for each meaningful phonological movement, creating extra keyframe

detections and making their alignment and modeling harder. On the other hand, in compound signs where gestures are composed of successive hand shapes, keyframe-based methods are the most beneficial. The results in Table 4.10 demonstrate that when combined with the 3D Convolutional ResNet approach from [158] through score level fusion, the keyframe-based TTN approach gives the state-of-the-art results on the BSign22k dataset. The results of score fusion also demonstrate that compounds and repetitive gestures are the signs that benefit most from keyframe-based TTN.

Table 4.10. The accuracies (%) of score fusion of the Conv3D MC3 method with Linear sampling, the keyframe-based method, TTN Alignment of Keyframes and TTN alignment of linear sampled frames are reported on different gesture subgroups: One-handed (1H), two-handed (2H), circular (C), non-circular (NC), repetitive (R), non-repetitive (NR), monomorphemic (MM), and compound (COM) signs.

	234 1H	510 2H	75 C	669 NC	457 R	287 NR	375 MM	369 COM	744 all
Conv3d MC3	90.87	**95.55**	**95.70**	93.85	93.37	95.08	92.21	95.96	94.03
Lin.Sampling	93.25	94.52	91.12	94.46	93.56	95.01	92.12	96.15	94.12
Keyframes	94.12	94.48	91.65	94.67	95.09	93.22	93.00	95.76	94.37
TTN Lin	**94.67**	94.06	94.22	94.25	93.61	**95.27**	**93.25**	95.27	94.25
TTN KF	93.90	95.36	91.29	**95.30**	**96.45**	92.43	92.65	**97.19**	**94.90**

4.1.2.9. Comparison with State-Of-The-Art Methods. In this section, the ATAF methods are compared against four similar coordinate based descriptors namely Hierarchical Co-occurrence Networks (HCN) [55], Sign language graph convolution networks (SL-GCN) and Separable Spatial Temporal Convolutional Networks (SSTCN) [16], and MHI+RGB from Mercanoglu *et al.* [91] on the BSign22k [10] and the AUTSL datasets [58]. Of the four methods, HCN and SL-GCN are methods that directly use coordiante values (i.e. 2 coordinate values for each joint of a 2D skeleton) for linearly sampled T time frames from J joints. On the other hand SSTCN, and MHI+RGB are heatmap based approach like ATAF. We present experimental results with each indi-

vidual method as well as each methods fusion with 3D-CNN methods. The experiment setup is shown in Figure 4.7. The results are presented in Table 4.11. In the table all results except MHI+RGB are reported by ourselves. Since [91] presents results on the same two datasets, we use their results directly from the paper.

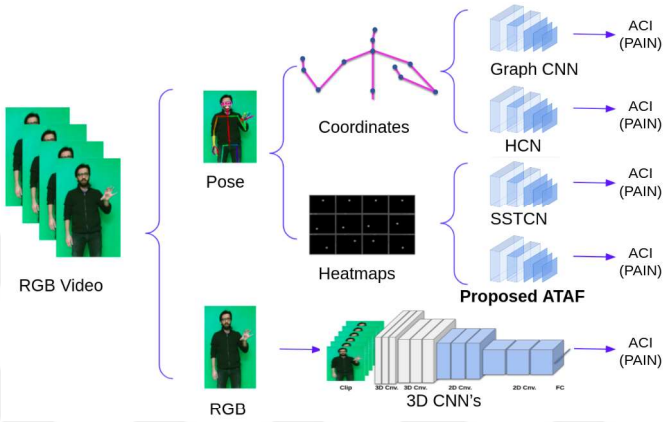

Figure 4.7. SLR Experiments on AUTSL and Bosphorussign22k Datasets.

As the dataset has fewer users and repetitions per class, learning is more challenging, and the accuracy scores are lower than the accuracy scores on AUTSL. Secondly, among the compared methods, the coordinate-based approach along with ATAF obtain the best accuracy in BSign22k while the heatmap-based and more complex alignment approach achieve better performance on the AUTSL dataset. This difference can be attributed to there not being enough samples to properly train the more complex heatmap-based alignment approach in an end-to-end fashion.

We observe that the performance of different methods relative to each other shows minor variations between the AUTSL and BSign22k datasets. Among the four skeleton-based methods, the graph convolution-based approach and the aligned ATAF approaches yield the best results both by themselves and also when their scores are fused with 3D-CNNs. The BSign22k dataset is a more challenging dataset as it contains three times as many classes and 10k fewer total videos.

Table 4.11. Comparison of ATAF using coordinate (ATAF-CD-KF) and heatmap (ATAF-HM-KF) based keyframe detection approaches with state-of-the-art heatmap based SSTCN and Motion History Image with RGB fusion and coordinate based Hierarchical Co-occurrence Network (HCN) and SL-GCN methods.

Method	Fusion	Dataset	
		BSign22k	AUTSL
HCN	-	35.12	31.77
GCN	-	81.10	89.12
SSTCN	-	77.81	90.78
ATAF-CD-KF	-	80.15	90.14
ATAF-HM-KF	-	79.44	92.36
MC3	-	86.25	85.33
GCN	MC-3D	**94.89**	97.01
SSTCN	MC-3D	91.60	95.11
[91] MHI+RGB	I3D	94.83	93.53
ATAF-CD-KF	MC-3D	**94.89**	97.09
ATAF-HM-KF	MC-3D	94.50	**97.59**

In table 4.12, we further observe the effect of clip size on each competitive method. Looking at the top methods in both datasets, we observe that keyframe-based approaches use fewer frames to obtain comparable accuracies. In our belief, models that use smaller clip sizes may suffer from information loss in discarded frames, while using longer clip sizes may lead to overfitting for the models we employ. Thus, for models

such as the ATAF feature that uses heatmaps to align keyframes, further improvement of the results presented in this tab may be possible using more complex models with larger clip and batch sizes.

Table 4.12. Analysis of clip length for the comparison of ATAF using coordinate-based keyframe(ATAF-KF), linear sampling(ATAF-LINEAR SAMPLING), and heatmap based keyframe detection (ATAF-HM-KF) approaches with state-of-the-art heatmap based SSTCN, HCN and GCN methods.

Model	Clip Length	Dataset	
		BSign22k	AUTSL
HCN	20	26.83	27.18
	40	31.10	30.20
	60	31.76	35.11
GCN	20	80.97	88.03
	40	81.06	89.12
	60	**81.10**	89.12
SSTCN	20	77.81	86.24
	40	75.42	90.78
	60	70.16	89.93
ATAF-KF	5	79.60	89.07
	7	**81.03**	90.78
	9	80.15	90.62
ATAF-Linear Sampling	20	78.03	89.28
	40	78.23	89.55
	60	77.81	88.91
ATAF-HM-KF	5	79.44	**92.36**
	7	79.24	92.25
	9	78.93	91.39

4.1.3. Discussion

In this chapter, we present the ATAF method, which is a pose-based visual representation built on the idea of aggregating joint heatmaps over isolated sign language videos. We propose a keyframe-based linguistic temporal colorization scheme that represents static frames in hand and finger movement to describe compound signs as a

collection of distinct subunits. To detect these keyframes, we implemented a static subunit detection method that makes use of local minima in hand speed with density clustering to extract dynamic and static subunits. We use an end-to-end alignment approach to align subunits so that extracted keyframes from differently performed gestures can be aligned with respect to speed and joint coordinate locations. We make use of additional features and joints, including detailed finger joint coordinates, and introduce an end-to-end training scheme by optimizing the ATAF features for fast and efficient extraction. An ablation study on feature construction is performed to find the best performing features as well as parameter optimizations for the use of different joint groups and model architectures.

In terms of aligning keyframes, we make use of end-to-end trained Temporal Transformers to align keyframes and linearly sampled frames. Analysis of keyframe extraction strategies with varying characteristics reveals the efficacy of the suggested method. The experiments demonstrated that compound gestures composed of two or more signs benefited from keyframe-based approaches in contrast to repetitive and circular gestures that benefited more from longer sequences of aligned uniformly sampled gestures. Our analysis demonstrated that modeling subunits with keyframes improved accuracy in gesture types that adhered well to the movement-hold model, such as non-circular, repetitive, and compound gestures, showing larger than average accuracy. In circular and mono-morphemic gestures, the accuracy obtained with features constructed by linearly sampled and aligned frames was considerably higher.

Sign language signs are complicated gestures with many degrees of freedom, where learning complicated representations from RGB images requires huge amounts of annotated data. Since the available sign language datasets are orders of magnitude smaller than the action recognition datasets both in variety and size, we demonstrated that it is possible to improve on the state-of-the-art in isolated SLR by combining state-of-the-art 3D-CNN methods with complementary features where joint coordinates of the person performing the sign language and the spatio-temporal subunits of the performed gestures are emphasized.

On the public AUTSL dataset, the method achieves 92.35% recognition accuracy, beating competitive skeleton-based approaches. Similarly, the method achieves a peak accuracy of 80.14% over 744 classes of the BSign22k dataset, which is higher than the 71.02% performance achieved with linear sampling and 51.2% achieved with the comparable Potion method [14]. The proposed method, which only uses joint coordinates, aims to serve a complementary role to that of 3D convolutional SLR approaches. Score fusion of ATAF with MC3 ResNet models [74] yields the top performance of 94.9% accuracy on the BSign22k and 97.6% accuracy on the AUTSL, surpassing top accuracy results obtained by methods using only the RGB image as input.

5. MULTI-TASK LEARNING APPROACHES

In this chapter, we detail our work on transfer learning and how we came to adapt our work on spatial transfer learning to the problem of sign language alignment. For this task we focused on two distinct approaches. First we focused on regularized Multi-task learning (MTL). Learning many related tasks simultaneously enhances modeling accuracy and feature selection, especially when each task has few training samples. Recognizing that not all activities are uniformly related, modeling task correlations in MTL is a popular approach. We chose personality prediction as a toy problem to implement and develop regularized regression based methods on a spatial a multi-modal prediction framework prior to moving our experiments to the temporal sign language videos. We explored the usage of several regularized regression based multi-task learning approaches]. Taking popular feature extraction / selection approaches such as PCA [20] and MRMR [21] for preprocessing and classifiers such as SVMs, Ridge regression and decision forests, we compared our transfer and multitask learning based personality prediction framework to classical machine learning approaches. In the next step, we moved our experiments to sign language recognition problem by using multiset alignment [22, 159] approaches to align sign language gestures and proposed a method called multi-task CTW by combining Canonical Time Warping framework with the spatial domain regularized multi-task learning approach we used in personality prediction. [19]

5.1. Regularized Multitask learning & Personality Prediction

The personality of a human is highly correlated with his behaviour and decisions in work and leisure environments. Through experience, humans can become quite good at judging the personality of their contacts and modulating their behaviour. Automated systems would benefit from the availability of an automated feedback mechanism based on the prediction of the personality of a user of the system.

Training personality predictors based on non-verbal audio-visual cues is a relatively new and scarcely explored territory in research. One major obstacle holding back research in this field is lack of large amounts of annotated data. Previous work on the subject used different approaches such as finding co-occurence features [160] or incorporating similar data from different contexts [161]. However the success of such systems often relies on jointly utilizing information from different sources and features as recognizing personality based on a low number of extracted features often do not yield high accuracy.

By their construction in psychology, the big five personality traits are supposed to be uncorrelated [162]. However, as Vinciarelli notes in [163], this is not often the case as raters' cognitive and cultural biases tend to create correlations among traits while annotating. In addition, personality traits that can be observed in group settings such as emergent leadership and dominance highly correlate with the big five personality traits, in particular with extraversion. To handle such correlation when building predictive models, we explore the usage of a Multi-task learning framework to search for a combination of tasks and features that highly correlate with the leadership and extraversion traits of humans.

In this section, we present a framework for recognizing extraversion and leadership from meeting videos of the ELEA corpus. We focus on recognizing these two traits, which has the highest correlation with our extracted nonverbal features. In previous studies, Sanchez-Cortes et al. [164] demonstrated the correlation of personality traits with the nonverbal features extracted on the ELEA corpus. In [161], we showed that a transfer learning approach leads to an increase in extraversion recognition performance, through the incorporation of another database (i.e. the VLOG database [165]) in the learning framework. Building on these previous work, we argue that the two traits, extraversion and leadership are related traits and a recognition framework for these traits could benefit from a joint learning approach.

5.1.1. Prediction of Personality Traits

Personality is defined as the combination of affects, behaviours, cognitions and desires that characterizes individuals in unique ways [166]. In the literature many approaches are used to model personalities of humans. The Big Five model proposed by psychology entails five dimensions of personality that have been used extensively in the literature to describe unique individuals [162]. The personality traits in the big five scheme are defined as follows:

- Agreeableness: friendly/compassionate vs. analytical/detached
- Conscientiousness: efficient/organized vs. easy-going/careless
- Emotional Stability: sensitive/nervous vs. secure/confident
- Extroversion: outgoing/energetic vs. solitary/reserved
- Openness to experience: inventive/curious vs. consistent/cautious

In addition to personalities defined for generic purposes, traits that are extensions of personalities such as emergent leadership, dominance and competence are significant where interaction between people are critical for success in undertaken tasks. Emergent leadership appears mostly in newly formed groups. The behaviour of a participant makes a person a good leader or not, without considering past information of competence, related task performance or friendship. This trait highly correlates with dominance and the big five personality traits [164]. In addition, Sanchez-Cortes et al. in [167] demonstrate that correlations exist between perception of leadership and automatically extracted visual cues.

Different methods are used to evaluate personalities of people the most common being questionnaires. The most widely used are NEO and TIPI questionnaires with 240 and 10 questions [168,169]. For tasks like automatic recognition, such questionnaire are used to generate ground truth annotations. These annotations are either generated by self-reporting of the user or the reporting of an observer. While self reporting presents a more accurate analysis of a persons personality, subjects tend to bias their answers

towards more socially acceptable results if the conducted test results may affect the person in a negative manner (i.e. failing a job interview) [169]. Therefore obtaining results from multiple observers is a common approach in evaluating personalities.

In the literature, the automatic recognition of personality from various multimedia based sources is a field of study that is gaining attention. One of the earliest studies on automatic personality prediction was done by Mairesse et al. [170]. The study explored the effectiveness of textual and speech based acoustic features in recognizing the big five traits. More recent studies on the topic focused on recognizing personality from a greater amount of data sources covering human interactions. These are text analysis based methods [170–172] focusing on choice of words, acoustic based methods focusing on non-verbal as well as verbal parts of speech [173, 174], wearable devices visual analysis based methods from image and videos [165, 175–177]. A more complete list of personality prediction literature can also be found in the survey of Vinciarelli et al. [163].

In automatic personality recognition research literature, the datasets and features pointing at personalities shows a large amount of diversity in terms of their data domains. These studies use the abundance of personal data from different domains such as short essays, blog posts [178], social media profiles [175], surveilance with wearable sensors [179, 180], computer game behaviour [181] and videos containing interaction between people [167, 174, 177].

One of the major efforts on automatic big five personality trait recognition occured with the Chalearn Looking at People challenge. The challenge organized by Lopez et al. [182, 183] had 10000 15-second job interview videos with continuous big five annotation scores in the range of $[0, 1]$. The challenge received many contributions using deep learning methods for feature extraction and personality prediction. The top group used a combination of CNN and LGBTOP based scene and face features with Extreme Learning Machines [184]. Other groups used multimodal LSTM Neural Network architectures to better represent the spatio-temporal nature of personality

recognition [185]. Some of the audio-visual features, first used in this study, were also utilized in this challenge and achieved scores that were comparatively close to these state of the art results with a simple single task random forest classifier [186].

5.1.2. Data and Annotations

We perform personality impression prediction on multi-person meeting videos of the publicly available ELEA (Emergent LEAder) corpus. The corpus includes 27 separate group meetings of 3 or 4 members [164]. This dataset includes audio and visual data and personality traits scored by external observers, such as the Big-Five personality traits, and by group members, such as the perceived leadership, dominance, competence, likeness, and ranked dominance.

In the videos of this dataset, the participants take part in a winter survival game. As the survivors of an airplane crash, the participants were asked to rank the importance of 12 items they would take with them to increase the chance of survival. After ranking the items individually, the participants ranked them as a group. A group of individuals are sat around a table, as they debate and try to persuade one another. Multi-microphone recording and wide-angle cameras capture the complete interaction scene.

The Ten Item Personality Inventory (TIPI), was used to gather a person's Big Five personality traits [168]. Each annotator watched a one-minute segment of a participant from the meeting, which corresponds to the segment that includes the participant's longest speaking turn. Three different annotators annotated each video and a total of five annotators annotated the whole dataset. In addition to the Big Five traits, the dataset also includes annotations for the perceived traits as annotated by the group members at the end of the meeting. These traits include perceived leadership, dominance, competence, liking, and ranked dominance. We will call these traits as the leadership traits in the rest of the paper. More details on the dataset and the annotations can be found in [187] and in [188].

In this paper, we used a subset from the ELEA corpus. The subset consists of audio-visual (AV) recordings of 27 meetings. The extracted videos of meetings contain long monologues by a single user. These segments contain parts of the meeting where the user is actively speaking.

In order to infer the personality and leadership traits, we extract different kinds of non-verbal audio-visual features as listed in Table 5.1.

Table 5.1. List of non-verbal features extracted from the ELEA dataset.

Feature ID	Features	Description
1:4	Speaking Status	Speech segments of the target person
5:8		Interruptions
9:18	Prosodic Speech Features	Pitch
19:28		Energy
29:32	Visual Activity	Head Motion
33:37		Body Activity
38:59	wMEI	Motion template based visual activity features
60:73	Visual Focus of Attention	Visual attention features

The second dataset we have used is the publicly available IDIAP VLOG dataset which contains the video blogs (vlogs) of video bloggers (vloggers) downloaded from YouTube. It contains a total of 404 videos. The dataset includes Big-Five personality annotations obtained by crowdsourcing. The personality annotations were collected via crowdsourcing on Amazon Mechanical Turk, using the TIPI questionnaire. Five different annotators annotated each vlog, after watching the first conversational minute of each vlog. More details about the dataset and the annotations can be found in [165]. Since it is recorded in a different social interaction setting, where vlogger is the only participant, extracting perceived leadership traits as in meeting videos is not possible.

The extracted features include visual features based on the subject's motion. The dataset contains a large number of samples for a small subset of features and tasks present in the ELEA corpus. The overlap of common tasks and features for the ELEA and VLOG datasets can be visualized in Figure 5.1.

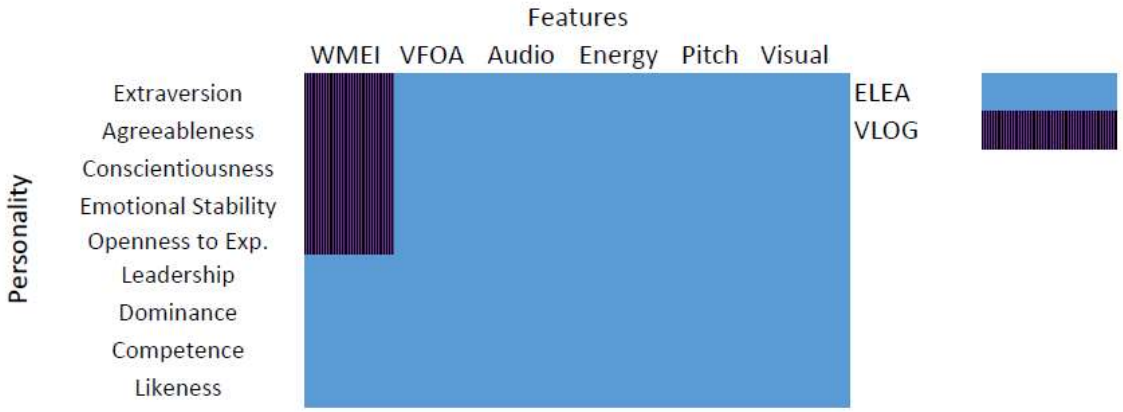

Figure 5.1. Set of personality traits and features used in prediction. VLOG segment displays overlapping set of features and tasks.

5.1.3. Methodology

In our experiments, we aim to perform recognition on the extraversion and leadership traits using the extracted audio-visual features from Table 5.1. For each subject in the dataset, the ground truth scores for the personality traits were obtained by averaging the personality scores coming from each annotator. For the ELEA dataset, each video is annotated for the Big Five personality traits by three different annotators. The leadership traits on the ELEA dataset were annotated by the fellow group members. It means either two or three annotators, excluding the subject, depending on the group size. For the VLOG dataset, each video is annotated for the Big Five personality traits by five different annotators. These traits were binarized by using the median values for each trait, thus creating a set of labels suitable for binary classification.

In the initial single-task prediction context, a task is defined as the learning of a personality trait. As there are two personality traits, we have two tasks. We consider each personality trait as a single task and classify them separately. For the baseline method, extraversion and leadership classification models were constructed using support vector machine, decision forest and ridge regression classifiers. Prediction accuracy for each task was calculated using a leave-one-out cross-validation approach.

We focus on the use of multi-task learning to improve extraversion and leadership trait predictions. Previous research on the ELEA dataset show that adding knowledge from other fields helps predict extraversion [189]. However, the lack of leadership-based annotations in VLOG, turns the problem into a heterogenous transfer learning problem, decreasing the transfer performance for leadership prediction. To tackle this difficulty, we use a multi-task learning framework to increase extraversion and leadership recognition performance. In the following sections, we explain our proposed feature selection, transfer learning and multi-task learning methodologies.

5.1.4. Feature Selection

The objective of feature selection is to eliminate noisy, irrelevant and redundant features to obtain a shorter feature vectors, where better classification is possible. This allows the building of better performing models with regard to both accuracy and speed while also presenting a better understanding for the data and the prediction process.

In the feature selection literature, there are two popular methods: filtering based and wrapper based methods [190]. Filtering based methods, which are more cost efficient, rely on extracting and comparing the inherent properties of features such as relevance, correlation, mutual information. Wrapper based methods on the other hand make use of black-box learning algorithms to predict learner performance using different search strategies.

In filtering based approaches, finding the most efficient subset of labels given a set of data and its class labels is possible through a method called maximum dependency. Using mutual information as a measure of similarity such an approach can calculate the best possible subset of features which have the largest dependency on the label of the given class. However such an approach requires the calculation of multivariate density functions for $p(x_1, ..., x_m)$ and $p(x_1, ..., x_m, c)$ which are difficult in high dimensional space.

In this study, we use Minimum Redundancy Maximum Relevance (MRMR), which is a frequently used feature selection algorithm. It is equivalent to the Max-Dependency algorithm for first order search. In MRMR, the mean value of all mutual information values across individual features and class labels are calculated as the Max Relevance values. The Minimal Redundancy between variables is calculated by finding the mutual information scores between variables. Finally, by using a function such as the difference of Max relevance and Min redundancy, the most dependent feature can be selected. Using this method in an incremental search, it is possible to find the m most important features.

5.1.5. Multi-Task Learning

Multi-task learning is a statistical spatial feature mapping framework where identical input data from different tasks are jointly learned. Joint learning of unrelated tasks can often lead to more informative sparse descriptors. Such descriptors can provide better discriminative power for the unusual features of the data. When the tasks are uncorrelated, their contribution can act as noise to other tasks thus improving generalization. Adding tasks may also increase the weight of parameters with more tasks allowing a better capture of the feature space.

In our formulation, we have a set of non-verbal features extracted from the ELEA corpus videos, represented as $X = x_{1..N}$, where each i represents the features from a single video sample. These are the same among tasks $t = 1..T$. The labels for each task

are denoted as $Y(t)$. The aim of our framework is to learn a dictionary W which maps the training samples, X to their labels, Y. Classical machine learning frameworks aim to learn a joint dictionary defined as $Y = WX$. Solutions to such problems can be obtained by using least square based methods. Desired properties of the dictionary can be obtained by adding a regularization term such as the element-wise $L1$ matrix norm as imposed by the LASSO algorithm [18]. As we can see from the LASSO algorithm represented as

$$\underset{W}{\operatorname{argmin}} ||WX - Y||_F + \alpha||W||_{1,1} \tag{5.1}$$

$$where ||W||_{p,q} = [\sum_{j=1}^{n}(\sum_{i=1}^{m}|a_{ij}|^p)^{q/p}]^{1/q}$$

there is no differentiation between inputs and regression coefficients for multiple tasks. Multi-task learning algorithms aim to minimize the sum of squared errors (SSE) for each task rather than finding an optimized SSE for all training samples, X.

5.1.5.1. Multi-Task LASSO. The Multi-Task LASSO(MTL) method [19] is an extension of the regular LASSO method proposed by Tibshirani et al. [18]. The multi-task LASSO makes it possible to fit multiple regression problems at the same time, making sure that the selected features are consistent between tasks. Formulated as

$$\underset{W}{\operatorname{argmin}} \sum_{i=1}^{t} ||W_i^T X - Y||_F + \alpha_1||W||_{1,1} + \alpha_2||W||_F \tag{5.2}$$

regularization parameters of MTL are α_1 and α_2, which control the sparsity of the dictionary matrix and the norm penalty respectively.

5.1.5.2. $L_{2,1}$ Norm Regularized Multi-Task LASSO. While MTL enforces a sparse representation on the entire dictionary, there is no emphasis on extracting the common features among the tasks. To enforce extracting shared features among tasks, Argyriou et al. [136], suggests adding a $L_{2,1}$ regularization term to the regularization problem. Since L_1 norm favours sparsity and the L_2 norm favours uniformity, formulating the problem as

$$\underset{W}{\operatorname{argmin}} \sum_{i=1}^{t} ||W_i^T X - Y||_F + \alpha_1||W||_{2,1} + \alpha_2||W||_F \tag{5.3}$$

where the $L_{2,1}$ provides the desired parameter sparsity shared across tasks by summing the euclidean norm of each column combines the best of both approaches.

5.1.5.3. Trace Norm Regularized Multi Task LASSO. Another multitask learning method making use of regularized regression is the trace norm regularized method proposed by Chen et al. [191]. The goal in trace norm regularized lasso is to find a low dimensional subspace shared by different tasks. In the regularization trace norm is used as a rank function.

The trace norm of the matrix $||W||_*$ is calculated by summing its eigenvalues. Minimizing the regression objective function in Equation 5.4 yields W matrices of minimal rank for each task as

$$\underset{W}{\operatorname{argmin}} \sum_{i=1}^{t} ||W_i^T X - Y||_F + \alpha_1 ||W||_* \tag{5.4}$$

where the parameter α_1 is used to control the rank of W.

5.1.5.4. Robust Multi-Task Learning. Robust Multi-Task Learning (ROBUST) algorithm focuses on the elimination of outlier tasks in the Multi-Task Lasso Framework. The model proposes that the model matrix W may be split into two matrices as

$$\underset{W}{\operatorname{argmin}} \sum_{i=1}^{t} ||W_i^T X_i - Y_i||_F + \alpha_1 ||L||_* + \alpha_2 ||S||_{1,2} \tag{5.5}$$
$$subject\,to : W = L + S$$

where each matrix represents task relevance and group sparseness.

In Equation 5.5, the trace norm enforces a low rank structure to couple tasks that are closely related, while the $L_{1,2}$ norm which is the row grouped L_1 norm captures outlier tasks.

5.1.5.5. Robust Multi-Task Feature Learning. Robust Multi-Task Feature Learning (RMTFL) algorithm focuses on the elimination of outlier tasks in the $L_{2,1}$ Norm Regularized Multi-Task LASSO framework [137]. In this framework, the W matrix is decomposed into two components as P and Q. The formulation of the method is given

as

$$\underset{W}{\operatorname{argmin}} \sum_{i=1}^{t} ||W_i^T X - Y||_F + \alpha_1 ||P||_{2,1} + \alpha_2 ||Q^T||_{2,1} \quad (5.6)$$
$$subject\,to : W = P + Q$$

where the P matrix captures group sparsity by jointly selecting sparse features across tasks while the Q matrix captures jointly selected sparse samples, capturing outlier tasks. The parameters α_1 and α_2 control the effect of each parameter in the learning process respectively.

5.1.5.6. Dirty Multi-Task LASSO. While the MTL and L_{21} Norm Regularized MTL performs well with ideal data, the data may not always be represented ideally using a single structure. Jalali et al, in [138], propose to decompose the model W into two components as P and Q, where one captures shared features among tasks while the other captures intrinsic properties that are useful in recognizing individual tasks. Formulated as

$$\underset{W}{\operatorname{argmin}} \sum_{i=1}^{t} \left\|W_i^T X - Y\right\|_F + \alpha_1 \left\|P\right\|_{1,\infty} + \alpha_2 \left\|Q\right\|_{1,1}$$
$$subject\,to : W = P + Q$$

where P captures group sparsity through $L_{1,\infty}$ norm which sums the maximum value of each row. Q enforces sparsity on the overall structure of the data and both P and Q are subject to W = P+Q.

5.1.5.7. Multi-Task Discriminant Analysis. Multi-Task Discriminant Analysis (MTDA) [192] can be seen as a multi-task extension of the widely used supervised dimensionality reduction technique Linear Discriminant Analysis (LDA) (Fukunaga 1991). Instead of simply pooling the data for multiple learning tasks together and learning a common transformation for all tasks, MTDA learns a separate transformation for each task.

Each MTDA transformation consists of two parts, one specific to the corresponding task and one common to all tasks. MTDA is based on an objective function, formulated as

$$\underset{W_i}{\operatorname{argmax}} \frac{tr(W_i^T S_b^i W_i M)}{tr(W_i^T S_t^i W_i M)} s.t. W_i^T W_i = I_{d'} \quad (5.7)$$

which is similar to that of the single task LDA. While most existing multi-task learning methods can only handle learning tasks with data sharing the same feature space, MTDA can deal with heterogeneous feature spaces, allowing the incorporation of data with missing features such as the VLOG dataset.

5.1.6. Transfer Learning

Many machine learning methods work well under certain assumptions such as having enough labelled data for each aspect of the features they model. In reality, such assumptions are usually not satisfied leading to imperfect models as is the case with the ELEA dataset [188]. The goal of transfer learning is to improve learning in a target task by utilizing knowledge from other source tasks.

In our previous study [161], the effect of transfer in personality affect prediction was analysed in ELEA and VLOG datasets. Baseline machine learning approaches were used in a transfer framework with target only, source only, combined and multi-task LDA based learning methods.

In this study, we employed different baseline and transfer based approaches for domain adaptation. Since the task of personality prediction used a shared set of features in both datasets, using multi-task learning methods to jointly learn from both datasets was possible. We made use of the regularized regression based multi-task learning approaches using 5 personality traits and 4 leadership traits from the ELEA dataset and 5 personality traits from the VLOG dataset as tasks.

5.2. Experiments and Results

We predict two personality traits, extraversion and leadership from meeting videos. The lack of sufficiently large amounts of annotated data necessitates the use of transfer and multi-task learning approaches to augment the feature space and to pre-process the data. Experimental results are presented for multi-task learning and

transfer learning frameworks. Various multi-task and transfer learning methods are used to augment the feature space and pre-process the input data. The need for both approaches arises from the lack of sufficiently large amounts of annotated data.

In our framework, multi-task learning approaches are employed to allow learning of multiple personality traits together in order to capture sparser and better representations on personality prediction models. On the other hand, Transfer learning approaches allow us to use features extracted from other domains.

In the baseline recognition system, the learning task is accomplished by Support Vector Machine [193], Decision Forest and Ridge Regression algorithms. LibSVM implementation is used for SVM and the C parameter of the SVM is selected from the $[2^{-5}, 2^{5}]$ range. Matlab's treebagger algorithm is used for performing learning with decision forests with 100 trees. Parameters for both methods were optimized using a two layered cross-validation scheme. In the ridge regression classifier, we used the 0/1 labels as the scores and estimated the label based on thresholding the estimated score at 0.5. Both source and target data were z-normalized, separately using each datasets' training samples. The ridge parameter values were selected from a range of $[2, 150]$

5.2.1. Transfer Learning Framework

In the transfer learning literature, domains are labeled as source and target domains according to their role in the learning process. Since we are interested in increasing the prediction accuracy of extraversion in small group settings, we choose the ELEA dataset as our target and the VLOG dataset as our source domains. We used the VLOG corpus to enhance a recognition model trained to recognize videos from the ELEA dataset.

Previous studies with the ELEA corpus using classical approaches obtained limited success due to a lack of sufficient training samples. As a method of easier bootstrapping, additional samples belonging to a similar domain can be utilized. In [194],

personality annotations in a video blog dataset were used to transfer information to the training models obtained from the ELEA dataset. However, while beneficial in allowing the creation of better models for extraversion, as a result of the differences between video blog and group meeting domains, transfer was only possible on the big five personalities and with visual WMEI features.

To investigate how target ELEA training data affected transferred recognition performance, we employed a Leave-one-out cross-validation technique in our trials. The target data, chosen as the leadership and extraversion traits of the ELEA dataset, is divided into 102 folds for 102 samples. Models using the 5 big five personality and 5 leadership annotations from the ELEA dataset and 5 big five personality annotations of the VLOG dataset are used as external data sources. Learning models are enhanced using the multi-task learning algorithms from section 5.2.2. We report the final accuracy on the ELEA dataset.

5.2.2. Multi Task Learning Framework

We used the samples of the ELEA dataset in a Leave-one-out cross-validation framework to obtain multi-task recognition results. In our experiment sets $D = \{X, Y\}$ we have 10 sets of binary labels Y for different tasks and only one set of features X extracted from one-minute video segments. The implementations of the methods are based on the MALSAR package [19]. In order to find the best-performing multi-task method alongside the best-performing learning algorithm, we attempt to find the best-performing feature/multi-task algorithm/learner combination with their different parameters. In our experiments, we use seven different multi-task approaches as summarized in Table 5.2.

These approaches are all utilized as feature extraction algorithms to obtain a new feature space representation. During training, we use a validation set to search for optimal parameters. In addition to searching for the best combination of learners and feature extraction methods, we also search for the multi-task and learning algorithm

parameters.

Table 5.2. List of Classifiers used in the study.

	NAME	ACRONYM
Regularized Regression Classifiers	Multi-Task Lasso	LASSO
	L21-Norm Regularized with Least Squares Loss	L21
	Trace-Norm Regularized Regression	TRACE
	Robust Multi-Task Learning	ROBUST
	Robust Multi-Task Feature Learning	RMTFL
	Dirty Multi-Task Learning	DIRTY
Multi Task Feature Learning	Multi-Task Discriminant Analysis	MTDA

5.2.3. Experiments

In our experiments, we test the prediction accuracies of different features, different classifiers, multi-task learning methods and transfer learning methodologies. We perform extensive tests and try to answer several research questions. In order to interpret the results better, we have organized our tests around the following questions:

5.2.3.1. Prediction of Individual Personality Traits. In the ELEA dataset, 10 personality traits have been annotated. In a meeting scenario, some traits are more strongly manifested than others. On the other hand, we make use of extracted audio and visual features to predict perceived personality traits. What will be the success of this prediction? To test this, we use all available features and use three popular non-deep learning-based classifiers: Random Forests (RF), Support Vector Machines

(SVM) and Ridge Regression (RIDGE). In Table 5.3, our tests illustrate that the prediction accuracy of extraversion and leadership are relatively high. Using the ridge regression method, the Leave-one-out cross-validation scheme demonstrates recognition accuracies of 75.5% for the extraversion trait and 68.6% for the leadership trait. It is observed that the recognition performances of the other personality traits are lower. Therefore we regard them as complementary tasks rather than focusing on their recognition. We predict only extraversion and leadership in further tests, and use the other tasks as complementary information in multi-task learning.

Table 5.3. Single task prediction accuracies for personality traits.

	Personality Traits				
	Ldr	Dom	Cmp	Lkn	RDom
SVM	64.7	68.6	57.8	55.9	60.8
RIDGE	68.6	53.9	54.9	64.7	55.9
RF	68.6	59.8	55.9	55.9	54.9
	Ext	Agr	Con	Emo	Opn
SVM	74.5	58.8	56.9	60.8	53.9
RIDGE	75.5	55.9	60.8	58.8	52.0
RF	73.5	52.0	51.0	50.0	52.9

5.2.3.2. How Useful Are the Individual Features. Next, we make use of the individual groups of features extracted from the ELEA dataset to analyze overall recognition performance of groups of features. We make use of three baseline two-class classifiers (SVM,RIDGE, RF), to evaluate the performance of each feature group as seen in Table 5.4.

In Table 5.4, bold figures indicate the best performance in each category. We observe that different features have varying predictive power for extraversion and leadership. Visual features perform well for the prediction of both traits. On the other

hand, WMEI features are more important for the prediction of extraversion while VFOA features are more important for the prediction of leadership.

Table 5.4. Prediction performance for individual feature groups.

	Extraversion			Leadership		
	SVM	RF	RIDGE	SVM	RF	RIDGE
Audio	46.1	59.8	54.9	55.9	59.8	55.9
Energy	59.8	62.7	66.7	53.9	57.8	62.7
Pitch	61.8	60.8	59.8	58.8	57.8	57.8
Visual	64.7	66.7	70.6	54.9	54.9	62.7
WMEI	60.8	65.7	62.7	56.9	55.9	61.8
VFOA	52.0	44.1	60.8	64.7	58.8	65.7
All	**74.5**	73.5	**75.5**	64.7	**68.6**	**68.6**

In order to concentrate on individual features rather than group of features, and test the effectiveness of their combinations, we employ feature selection (MRMR) and extraction (PCA) for single task prediction. The results are presented in Table 5.5. Baseline results with all features are given in the first row for comparison. We observe that feature extraction (PCA) is not beneficial; and decrease the recognition accuracy. Feature selection, on the other hand, increases performance. It is observed that the peak performances for extraversion prediction are obtained at a subset size of 40 and 50 features using MRMR and SVM's and 50 features with RDF's. For leadership prediction, the peak results are given using MRMR with random decision forests using 30 and 40 features while Ridge regression achieves a good performance of 71.6% using only 10 features.

Table 5.5. Effects of feature selection and extraction approaches for single task prediction for extraversion and leadership.

		Extraversion			Leadership		
	# Ftrs	SVM	RIDGE	RF	SVM	RIDGE	RF
None	73	74.5	**75.5**	72.5	64.7	68.6	67.6
MRMR	10	72.5	71.6	75.5	64.7	**716**	696
	20	72.5	67.6	75.5	57.8	64.7	71.6
	30	72.5	66.7	74.5	57.8	66.7	**72.5**
	40	**76.5**	72.5	75.5	**67.6**	63.7	**72.5**
	50	**76.5**	70.6	**76.5**	**67.6**	58.8	65.7
	60	75.5	73.5	74.5	66.7	65.7	71.6
	70	74.5	73.5	74.5	65.7	69.6	69.6
PCA	10	57.8	66.7	58.8	52.0	64.7	61.8
	20	67.6	72.5	64.7	61.8	63.7	57.8
	30	69.6	67.6	65.7	62.7	58.8	57.8
	40	70.6	74.5	68.6	60.8	60.8	63.7
	50	74.5	70.6	66.7	62.7	66.7	55.9
	60	73.5	69.6	62.7	65.7	64.7	56.9
	70	73.5	62.7	66.7	65.7	57.8	55.9

Exploration of the selected features for each trait demonstrates that while features for leadership prediction trait are mostly selected from speaking status and visual focus of attention features, features for extraversion prediction are selected from nearly all the available feature sources.

5.2.3.3. How Beneficial Is Multi-Task Learning. As described in Section 2, there are correlations in perceptions of personality traits; and this leads to correlations in the annotated data. This leads us to hypothesise that we can increase the prediction accuracy of extraversion and leadership using other personality traits. To test this hypothesis, we have conducted experiments on the ELEA corpus as summarised in Table 6.

We use 5 personality and 5 leadership traits as different tasks to jointly learn prediction models and use them with 9 different classifiers. Since only leadership and extraversion traits correlate highly with the features we extracted in the ELEA dataset,

we observe and report prediction performances only on those two traits.

We further include experiments with MRMR and MTDA feature selection approaches to search for the best performing learner combination. We observe that the best accuracies are 78.4% for extraversion and 72.6% for leadership, displaying a 2% increase from the baseline results. The results demonstrate that in contrary to the baseline methods, the regularized regression based methods do not benefit from the utilized feature selection and preprocessing approaches. This can be explained away by the fact that contary to single task methods, these classifiers have the ability to utilize features that are only useful for some tasks and are detrimental to others. The best results are obtained using the multitask Lasso method, which solely focuses on sparse features and does not take into account whether they are shared across tasks.

Table 5.6. Multi-task recognition for leadership and extraversion traits by utilizing all personality and leadership traits on ELEA.

Preprocessing:		None		MRMR-45		MTDA	
		Lead	Ext	Lead	Ext	Lead	Ext
Baseline	RF	67.6	74.5	**68.6**	**75.5**	61.8	72.5
	SVM	64.7	74.5	67.6	**75.5**	62.7	**75.5**
	RIDGE	68.6	75.5	58.8	70.6	65.6	70.6
Regularized Regression	LASSO	**70.6**	**76.5**	**68.6**	74.5	61.8	74.5
	L21	66.7	72.5	59.8	72.5	64.7	73.5
	TRACE	69.6	72.5	59.8	72.5	**65.7**	71.6
	ROBUST	68.3	72.5	58.8	69.6	56.9	69.6
	RMTFL	69.6	72.5	59.8	69.6	64.7	72.5
	DIRTY	66.3	76.5	59.8	73.5	61.8	73.5

5.2.3.4. Can We Transfer Knowledge from a Richer Domain. We next evaluate the performance of transfer learning. For transfer learning, we use VLOG as the source

domain and ELEA as the target domain. The VLOG dataset contains only a subset of the features in ELEA. The common features are the WMEI features. Therefore, only traits that are well represented by WMEI are positively affected by the transfer. Therefore, we focus only on extraversion prediction. The effect of using different methods for transfer learning for extraversion prediction is illustrated in Figure 5.2. We observe that Lasso and Dirty Regularization based Regression methods yield significant improvements in performance, both yielding 78.4% prediction accuracy.

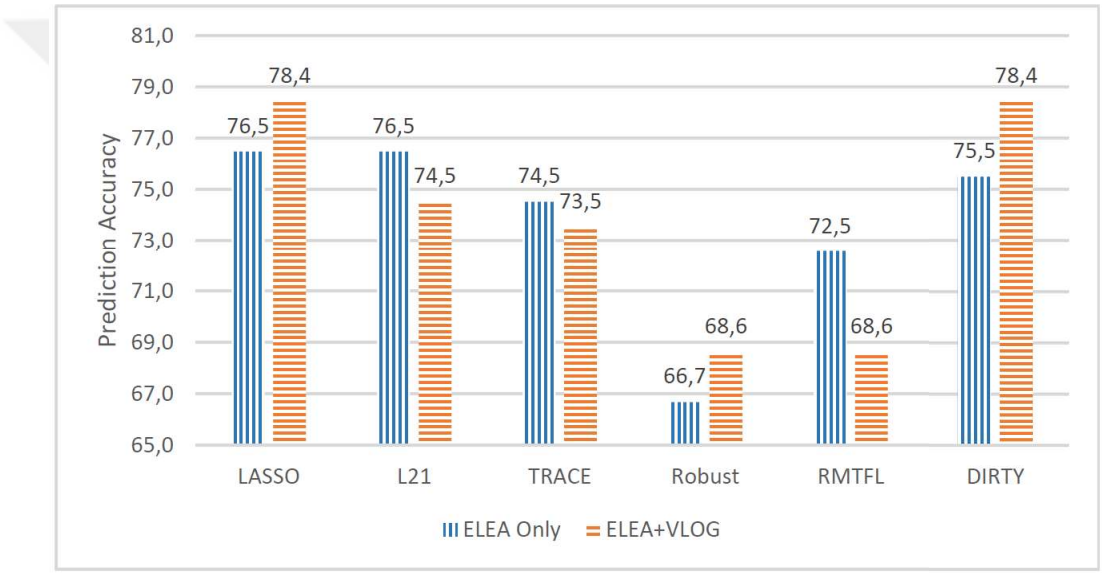

Figure 5.2. Prediction Accuracy of Transfer Learning in Extraversion Prediction.

5.2.3.5. Is Knowing Too Much Detrimental. In multi-domain learning, negative transfer across tasks is a problem. Trying to learn uncorrelated tasks together may often cause a decrease in performance. Therefore, better performances may be obtained by transferring only from relevant domains. We explore combining the best performing tasks in regularized regression based learning algorithms. For this purpose, we make use of forward task selection and incorparate the most relevant tasks one by one. Of the 15 tasks evaluated with forward selection, 5 are big five personality traits, 5 are emergent leadership traits from the ELEA dataset and the remaining 5 are big five personality traits from the VLOG dataset.

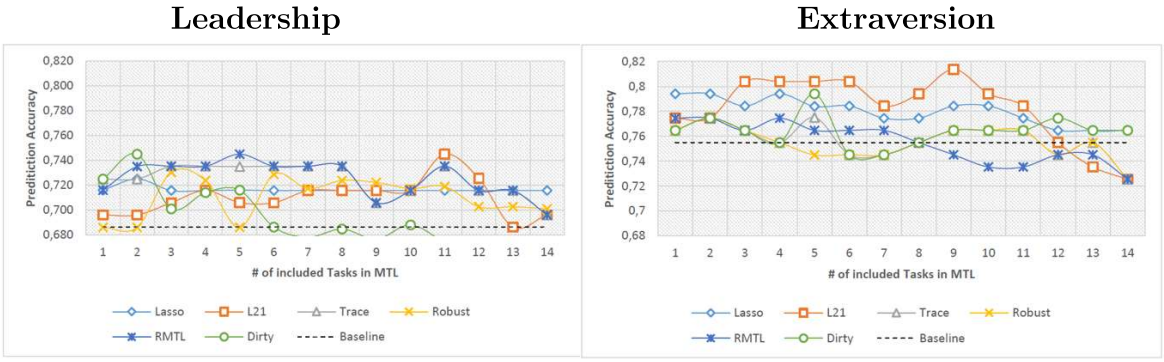

Figure 5.3. Multi-Task learning using forward selection of tasks. At each step, the best performing personality trait is included in the training set. Prediction performance on ELEA dataset for the leadership trait (left) and the extraversion trait (right).

As seen in Figure 3, incorporation of the most correlated tasks at each step of forward selection yields significant increases in the prediction of extraversion and leadership. This increase in performance is more evident in certain methods: L21 shows the best overall performance in both extraversion and leadership prediction. This leads us to search for the tasks that would yield the highest recognition performances when learned together. We observe that among the leadership prediction methods, Dirty MTL reaches top leadership prediction performance using 2 tasks. However, as the number of tasks increases, its performance drops. On the other hand, the recognition performance of L21 and methods that can deal with outliers better such as RMTFL reach top recognition performance. For extraversion prediction, the best results are given by L21 norm regularized multi task lasso method. However, L21 method performs worse than Multi-task Lasso when the task count is low and worse than DIRTY MTL when all the tasks are included.

For the trait of leadership, the most beneficial and therefore first incorporated traits in the forward selection scheme were Composure, Agreeableness, Emotional Stability from the ELEA and Openness to Experiences from the VLOG datasets in the given order. On the other hand, for the extraversion trait, Extraversion and Conscientiousness from the VLOG domain were selected first, followed by the dominance and likeness traits of ELEA. These results hint that while the largest increases in recognition performance for extraversion come from the VLOG domain data, recognition performance for leadership only benefits from utilizing other tasks of the ELEA corpus.

5.2.3.6. What Is the Best Overall Strategy. Table 7 summarizes the previous work on extraversion and leadership prediction on the ELEA dataset and compares them with the results in this paper. In extraversion prediction, the best results were obtained as 72.5% using WMEI features with transfer learning [161] and as 74.5% using the full set of features used in this study with baseline classifiers [188]. Our baseline classifiers exceeded this by 2% using feature selection approaches. We have obtained almost 6% improvement by using L21 regularized MTL with task selection. Adding transfer learning by itself also leads to about 3% improvement over the baseline. In Leadership prediction, the highest results on the ELEA dataset using a similar baseline methodology was reported as 74.5% in [160] using co-occurence features. Since our features do not include co-occurence, this result is not directly comparable with ours. In this trait, our baseline classifiers achieved a top recognition performance of 68.6%. Using multitask learning with task selection, we were able to obtain 74.5% performance, reaching the top performance for the leadership trait on the ELEA dataset that was reported with co-occurence features.

One interesting result that can be observed is the performance drop of leadership prediction for the MTL method with multi-task LASSO (70.6%) compared to the single-task Ridge Regression classifier baseline with MRMR feature selection (72.5%). From this, it can be deduced that when using all the tasks, learning with a subset of features is better than using all. However, when the most beneficial tasks are chosen,

building regression based classifiers using all the features increases peak performance.

Table 5.7. Performance on the ELEA Dataset.

	ELEA Dataset Performance	
	Extraversion	Leadership
[161]	72.5	-
[188]	74.5	-
[160]	69.6	**74.5**
Baseline - Ridge	75.5	68.6
Feature Selection & RF	76.5	72.5
MTL - LASSO	76.5	70.6
Transfer - Lasso & Dirty	78.4	-
Forward Selection - L21	**81.3**	**74.5**

While trying to learn all 10 tasks from the ELEA and 5 tasks from the VLOG dataset did not result in an observable increase in prediction accuracies, a forward selection scheme for choosing the best tasks to learn jointly yielded best performances for the extraversion and leadership traits. Our learning methodology captures the shared information in different personality traits using the two compatible ELEA and VLOG datasets.

In conjunction with data transfer of personality traits from the video blog domain, we obtain 74.5% accuracy in leadership and 81.3% accuracy in the extraversion personality traits, exceeding the performances reported in [160] by ∼13% and in [161] by ∼9%. The effectiveness of the L21 norm based regularization hints us that detecting commonly used features across with other personality annotations is a critical aspect

in multi-task personality detection in meeting videos. Overall, this indicates that while applying transfer learning and learning multiple traits separately improve automatic leadership and extraversion recognition accuracy, their combination yields a greater increase in creating a better recognition framework.

6. TASK AWARE CANONICAL TIME WARPING

In sign language recognition (SLR), well-known methods such as 3D Convolutional Neural Network (3DCNN)s [74] and Multi Stream Networks [195] excel at locating differences between signs in the user's sign space during a brief time interval. However, due to the phonetic nature of Sign Languages, a significant percentage of the calculation is wasted searching for spatio-temporal intervals that are unlikely to contain meaningful information. In situations where there are insufficient data per class to adequately train spatio-temporal models. These models may fail in straightforward tasks, such as comparing similar parts of signs or subsequence matching. In such situations, we expect that using temporal mechanisms capable of including SL domain information will improve the recognition performance of methods that incorporate temporal sampling, such as random forest-based approaches and CNNs with temporal pooling. In [36], Lidell et. al. suggested a grammatical model for sign language gestures in which the signs are formed of movement and hold components. Researchers have applied this model to the problem of sign language recognition, as focusing on learning handshapes and upper body configurations during the static parts while focusing on movement trajectories during the dynamic parts is useful for distinguishing meaningful signs from involuntary or personal hand, body, and face movements.



Figure 6.1. Existing indirect alignment approaches vs proposed direct alignment approach.

In this section, we perform direct temporal registration of large number of different gestures using features such as hand speed, acceleration, hand trajectory orientation and hand shape. The method differs from the majority of the methods that perform feature extraction followed by indirect alignment and classification as seen in Figure 6.1. Using multiset alignment approaches such as DTW, CTW [22], GTW [23] and the proposed multi-task CTW, common dataset specific alignments are determined and the mappings of each gesture to that alignment is computed. Coupled with different spatial and temporal learning strategies, the efficiency of these methods are explored.

6.0.1. Temporal Registration For Sign Language Recognition

In sequential pattern recognition, machine learning methods need to cope with temporal variances as well as spatial differences. Several studies demonstrate that [154, 196] temporal alignment prior to classification allows the use of less complex learners thereby both reducing noise and increasing training speed on larger datasets. For this reason, we explore the use of Dynamic Time Warping, Canonical Time Warping and Task Aware Canonical Time warping based temporal alignment and keyframe selection strategies to align gesture sequences and subsequences prior to classification.

The task aware CTW method we propose builds upon the works of Zhou and De La Torre [159, 197]. Combining dynamic time warping with canonical correlation analysis-based feature selection, the CTW approach performs temporal registration on an uncorrelated subset of the database's features. The least square formulation given as

$$\begin{aligned} X &= [x_1, \cdots, x_{n_x}] \in \mathbb{R}^{d \times n_x} \\ Y &= [y_1, \cdots, y_{n_y}] \in \mathbb{R}^{d \times n_y} \end{aligned} \tag{6.1}$$

where X and Y are two time series.

The algorithm stretches the data in the temporal domain by replicating the columns of X to fit Y as can be seen in Figure 6.2 and with the equation given as

$$\min_{p_x,p_y\in\Psi} J_{dtw} = \|XV_x - YV_y\|_F^2 \tag{6.2}$$

$$where\, V(p) : \{1:n\} \rightarrow \{0,1\}^{n\times l} \tag{6.3}$$

where V is the warping matrix. In order to extend the least squares DTW formulation to aligning more than 2 sequences, procrustes analysis(Wiley, 1998) is used. Given m sequences the algorithm seeks warping paths $\{p_i\}_i$ to obtain alignments to the mean of the set of sequences such that

$$\begin{aligned} \min_{\{p_i\in\Psi\}_i} J_{pdtw} &= \sum_{i=1}^{m}\sum_{j=1}^{m}\frac{1}{2}\|X_iV_i - X_jV_j\|_F^2 \\ \min_{\{p_i\in\Psi\}_i} J_{pdtw} &= m\sum_{i=1}^{m}\|X_iV_i - \frac{1}{m}\sum_{j=1}^{m}X_jV_j\|_F^2 \end{aligned} \tag{6.4}$$

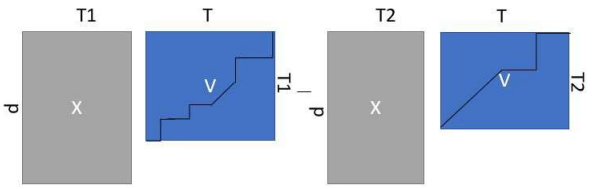

Figure 6.2. Warping sequences with DTW.

The procrustes or multi-set dynamic time warping algorithm computes a mean gesture value and then finds the best alignment matrix for each gesture that minimizes their DTW distance to the common mean value iteratively.

In [197], the authors extends the alignment of DTW by simultaneously performing canonical correlation analysis (CCA) and temporal alignment making sure that only features that are correlated over samples are used in the alignment process. By offering a feature selection and dimensionality reduction method for aligning signals on various dimensions, the CTW method provides more reliable alignment. Combination of the CCA algorithm with the least square DTW formulation can then be given as seen in

Figure 6.3 and formulated as

$$\min_{\{V_x,V_y\}\in\Phi,\{p_x,p_y\}\in\Psi} J_{ctw} = \|V_x^T X W_x - V_y^T Y W_y\|_F^2 + \phi(V_x) + \phi(V_y) \tag{6.5}$$

$$where \quad V_x \in \mathbb{R}^{d_x \times d} \quad and \quad V_y \in \mathbb{R}^{d_y \times d} \tag{6.6}$$

$$\Phi = \big\{\{V_x, V_y\} \mid V_x^T\big((1-\lambda)XW_xW_x^TX^T + \lambda I\big)V_x = I \\ V_y^T\big((1-\lambda)YW_yW_y^TY^T + \lambda I\big)V_y = I\big\} \tag{6.7}$$

$$\begin{bmatrix} 0 & XV_X^TV_YY^T \\ YV_XV_Y^TX^T & 0 \end{bmatrix} W = \begin{bmatrix} XV_X^TV_XX^T & 0 \\ 0 & YV_Y^TV_YY^T \end{bmatrix} W\Lambda. \tag{6.8}$$

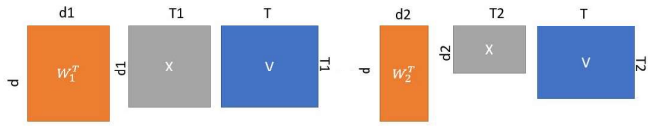

Figure 6.3. Finding correlated features with CTW.

The sequence alignment of CTW is initialized through the use of uniform scaling or some alternative method which brings the gestures to the same length [198]. The iterative solution of the CTW method is obtained through the algorithm in Figure 6.4.

6.0.1.1. Generalized Canonical Time Warping. This method was proposed by Zhou et al. [23] to extend CTW by replacing the alignment calculation with dynamic programming part by a flexible path warping as a combination of several monotonic basis functions. For a given collection of m time series

$$\{X_i\}_{i=1}^m \\ X_i = [x_1^i, \cdots, x_{n_i}^i] \in \mathbb{R}^{d_i \times n_i} \tag{6.9}$$

GCTW tries to find for each X_i, a spatial embedding of low dimension

$$W_i \in \mathbb{R}^{d_i \times d} \tag{6.10}$$

Input: X, Y

Output: V_x, V_y, W_x, W_y

Initialize $W_x = I_d x, W_y = I_d y$

while While J_{ctw} is not converged **do**

Utilize dynamic programming to compute V_x, V_y for sequence alignment $W_x^T X, W_y^T Y$

Set the leading b generalized eigenvectors of Equation 6.8 to the columns of $W^T = [W_x^T, W_y^T]$

end while

Figure 6.4. Canonical time warping algorithm.

and a non-linear temporal transformation

$$V_i = W(p_i) : \{1 : n_i\}^l \rightarrow \{0, 1\}^{n_i \times l} \tag{6.11}$$

such that,

$$W_i^T V_i W_i \in \mathbb{R}^{d \times l} \tag{6.12}$$

is well aligned with others in a lest-square comparison. GCTW minimizes the sum of pairwise sequences distances:

$$\begin{aligned} \min_{\{W_i\}_i \in \Phi, \{p_i\}_i \in \Psi} J_{gctw} &= \sum_{i=1}^{m} \sum_{j=1}^{m} \frac{1}{2} \|W_i^T X_i V_i - W_j^T X_j V_j\|_F^2 \\ &+ \sum_{i=1}^{m} \big(\phi(W_i) + \psi(p_i)\big) \end{aligned} \tag{6.13}$$

$$\begin{aligned} \phi(W_i) &= \frac{m\lambda}{1-\lambda} \|W_i\|_F^2 \\ \lambda &\in [0, 1] \end{aligned} \tag{6.14}$$

$$\Phi = \big\{ \{W_i\}_i \mid \sum_{i=1}^{m} V_i^T \big((1-\lambda) X_i V_i V_i^T X_i^T + \lambda I\big) W_i = I \big\} \tag{6.15}$$

The time warping function is optimized using a linear complexity *Gauss-Newton* method. Its details can be explored at [23].

6.0.1.2. Clustering based Task Aware CTW. In CTW, the correlation search algorithm searches for shared correlation among all samples. Since the amount of diversity

in our dataset is quite large, searching for correlation among groups of samples may actually lead to better alignments. To find such groupings, we borrow from the multi-task learning literature and use spectral clustering similar to that of [23] Zhou et al.'s Clustered Multi task learning approach.

In this approach, we modify the canonical time warping algorithm by grouping the alignment matrices V and canonical correlation matrices W and solving them independently with each task as can be seen in Algorithm 6.5.

Input: X, Y, k

Output: V_x, V_y, W_x, W_y

Initialize $W_x = I_d x, W_y = I_d y$

while While J_{ctw} is not converged **do**

Compute $J_{i,j}$ using V and W matrices

Perform Spectral Clustering using $J_{i,j}$ as an adjacency matrix

For samples belonging to each cluster K, Use dynamic programming to compute V_x, V_y for aligning the sequences $W_x^T X, W_y^T Y$

For samples belonging to each cluster K, Set columns of $W^T = [W_x^T, W_y^T]$ to the leading b generalized eigenvectors of Equation 6.8

end while

Figure 6.5. Task Aware Canonical Time Warping.

6.0.1.3. Keyframe Selection. The second stage of the proposed temporal registration approach is the selection of keyframes. In Sign Language recognition literature, stationary frames where hand shapes and locations are meaningful are designated as keyframes. In the Lidell movement-hold approach keyframes are the static gestures between moving frames. These frames are sufficiently informative to describe a majority of the gestures in isolated SLR datasets. In this study, representative frames for

recognizing signs are extracted using three methods:

- Linearly Spaced Samples: Choose a fixed number of consecutive frames with equal intervals to represent the entire video. Selecting the interval as the total number of frames allows all frames to be selected as keyframes.
- Aligned Linearly Spaced Samples: The frames of the selected videos are aligned with respect to each other prior to sampling representitive frames. When the gestures are aligned, the frames with the same temporal indices are chosen from all gestures.
- Keyframe selection: A keyframe selection using changepoints in hand speed and orientatiton is applied to reduce the number of possible frames before the alignment approach is applied to the sign.

6.0.2. Sign Language Recognition

Following temporal alignment, we utilize a sign language recognition methodology that makes use of two different learners with classifiying temporal image volumes. These are random decision forests with OpenPose Library based features [65] and 3-Dimensional Convolutional Networks [199]. The random forest based method makes use of several features constructed from features extracted from the openpose library while the C3D features are extracted directly from temporal rgb volumes.

The features extracted from openpose include the Hand speed feature, obtained by calculating the movement of the hand joint in 3D coordinate space; the hand joint distances, extracted by calculating the euclidean distance between the hand joints and the upper body joints, namely shoulders, elbows neck and head; and the hand shape cnn features explained in the work of Ozdemir [200]. We explored the use of several discriminant based classifiers to recognize gestures with these features and found Random decision forests to be the spatial classifier of choice along with Subspace discriminant based linear classifiers.

We use the 3D Convolutional Networks proposed by Tran *et al.* [199] and use the pretrained weights from their model to initialize our network. 3D CNN's utilize both temporal and spatial information from consecutive frames to perform learning on videos. Their main advantage to 2D CNN's is that they remove the need for temporal modelling. 3D Convolutions are computed from a volume and temporal dependencies are modeled through training a model where desired spatio-temporal activation's are rewarded.

In Sign Language Recognition using C3D methods, each gesture is represented by a single volume constructed of 16 frames. In [199], the authors propose the use of non-overlapping windows of length 16 to represent a video for action recognition, while in [201], Camgoz et.al proposes the use of overlapping windows of length 16 to represent each frame. These temporal windows are then majority filtered and averaged to make gesture-wise decisions. Unlike the original proposal of the algorithm, we have explored that selecting a fixed number of 16 frames from a video using temporal registration and using those frames to make a classification decision for the entire gesture works better and is more efficient in our dataset of isolated gestures.

6.0.3. Experiments

In this experiment, two subsets of the Bosphorus Sign Dataset are used. The Overlapping Subsequences Dataset(OSD) contains 400 instances of 10 gestures from 5 users. The Bosphorus Sign - General Dataset is larger containing 5829 gestures from 154 Sign Language Classes. The gestures are segmented, but each gesture may be composed of more than one gloss meaning that each video may share several distinct subsequences with other videos. One example of such a pair from the OSD are the gestures "emergency" and "is this an emergency?" where the former gesture has additional manual and facial components following the components of the first gloss. The experiments performed on the dataset were all performed using the overall hardest user (User 4) as a test set, ensuring that all performed tests were user independent tests. Leave one user out tests were performed with the Random Decision Forest based

method to further validate results.

6.0.3.1. Baseline Methods for Classifying Signs. In this first experiment, we compare the performances of several state of the art methods for classifying gestures on the Bosphorus Sign Dataset. The experiments are performed on the smaller Bosphorus sign OSD dataset of 10 isolated gesture classes. We compare recognition performances of Random Decision Forests, Two Stream Convolutional Neural Networks, Bi-Directional LSTM's with VGG features and Convolutional 3D Networks.

Table 6.1. Baseline classifier accuracies for BS-SSS Dataset.

	Sliding Temporal Window	Linearly Spaced Samples
RDF	0.762	0.805
Two-Stream	-	0.72
VGG+LSTM	-	0.68
C3D Networks	0.746	0.76

Based on the results presented, we found that LSTMS and Two stream networks displayed a sub-optimal performance in this problem. For that reason, we continued the remainder of these experiments with RDF and C3D Networks.

6.0.3.2. Temporal Registration of Signs Prior to Classification. In this experiment, we evaluate several methods using hand speed to pre-align videos before performing classification. RDF and 3D Convolutional Neural Network classifiers are used in the experiments to observe performance. We compare the methods of procrustes Dynamic Time Warping (pdtw), procrustes Canonical Time Warping and generalized Canonical Time Warping. The results shown in Table 6.2 demonstrate the results.

Table 6.2. Analysis of the alignment methods used for temporal registration.

	OSD		General	
	RDF	3D Conv. NN	RDF	3D Conv. NN
Lin. Int.	0.805	0.760	0.349	0.553
PDTW	0.808	0.958	0.395	0.571
PCTW	0.879	0.958	0.376	0.524
PGTW	0.858	0.948	0.364	0.551

The results present in Table 6.2 demonstrate an increase in gesture recognition performance when the results are temporally aligned to a mean gesture of the training dataset.

Table 6.3. Exploring the effect of task wise clustering on canonical time warping.

		RDF	Subspace LD
	# of Clusters		
Spectral Clustering - only SSE	3	0.910	0.9158
	5	0.9131	0.9237
	7	0.9284	0.9131
	10	0.9053	0.9105
Spectral Clustering Regularized SSE	3	0.9237	0.9105
	5	0.9368	0.9237
	7	0.9237	0.9157
	10	0.9105	0.9105

6.0.3.3. Negative Transfer Avoidance by Clustering Sequences. In the experiments so far, the entire dataset was treated as a single entity where we aligned to find a common mean sequence and align each sequence to that. However, we hypothesized

that due to the amount of large variance in the dataset, this may actually reduce performance. To remedy this, we utilize the use of task aware canonical temporal alignment with different cluster sizes and analyze the results.

Table 6.4. Exploring combined the effect of user wise grouping combined with task wise clustering on canonical time warping.

		Number of Sub-Clusters		
		1	2	3
# of User Groups	1	0.8974	0.9079	0.9079
	2	0.8947	0.9105	0.9105
	3	0.8868	0.9000	0.9158
	4	0.8868	0.8974	0.9105
	5	0.8895	0.8895	0.8895
	6	0.8895	0.8816	0.8842

6.0.3.4. Analysis of Temporal Strategies for C3D Features. In this experiment we have evaluated several temporal selection strategies for C3D as suggested in the literature and compared them with our alignment approaches to find better classification results.

We have explored the usage of averaging overlapping and non-overlapping sliding windows to obtain a sign-wise classification decision. Similarly We have used linearly sampled equidistant variable length frames (linspace), randomly sampled frames from linearly spaced sampling intervals, and keyframes.

Deriving from the movement-hold model of linguistically modeling sign language gestures, in our experiments, we used three different features to align gestures. These are hand speed, hand distance to joints and hand shape information. These features while satisfactory on their own contain complementary information. For that reason, we evaluate their alone and combined performance with our alignment approaches.

Table 6.5. Evaluation of SLR accuracy with linearly spaced samples temporal strategy.

		FS1	FS2	FS3
OSD	pCTW	0.958	0.87	0.94
	GTW	0.948	0.88	0.95
	tCTW	0.958	0.88	0.91
GENERAL	pCTW	0.521	0.6534	0.6534
	GTW	0.551	0.6873	0.6892
	tCTW	**0.556**	0.5432	0.5432

Table 6.6. Evaluation of SLR accuracy with random samples from linearly spaced intervals strategy.

		FS1	FS2	FS3
OSD	pCTW	0.90	0.84	0.84
	GTW	0.99	0.98	0.98
	tCTW	0.99	0.98	0.98
GENERAL	pCTW	0.6882	0.6978	0.7035
	GTW	0.6666		
	tCTW	**0.6935**	0.7035	0.7005

6.0.4. Discussion

In this study, we have explored the usage of Canonical Correlation based temporal alignment methods to perform temporal registration on Sign Language Gestures. The main aim of the study was to explore the usage of temporal registration to improve recognition performance with classical approaches such as RDF's and 3D convolutional neural networks. Our experiments have demonstrated that for the problem of isolated

sign language recognition from videos, temporally registering gestures prior to classification increases recognition accuracy compared to brute force approaches such as using temporal sliding windows.

Table 6.7. Evaluation of SLR accuracy with keyframe based linearly spaced intervals strategy.

		FS1
OSD	pCTW	0.90
	GTW	0.99
	tCTW	0.99
GENERAL	pCTW	0.55
	GTW	0.58
	tCTW	0.58

In addition, the task aware CTW method we presented further increases performance by clustering the gestures in the dataset into subgroups independent of class or user information. By aligning similar gestures such as longer gestures, shorter gestures or gestures with different number of movement and hold segments together, the proposed method finds common groupings among the samples of the training set to make more natural alignments.

7. DEEP DOMAIN ADAPTATION FOR SIGN RECOGNITION

In this chapter, we move on to performing transfer learning for sign language recognition. Sign Language recognition is a video classification problem. State of the art video classification methods require huge datasets to be trained on. Some examples include Kinetics [25], and MomentsInTime [26] with thousands of video samples per class. In comparison to these datasets, the annotated resources available for isolated sign language recognition are smaller. In addition, even if in the near future large sign language datasets are collected for one or two major languages, there will be some under-resourced sign languages or sign language dialects that will not have access to large amount of samples per class. For this reason, in order to train models that are invariant to changes such as pose, dataset, signer and environment, transfer of learned knowledge from some other domains is necessary.

When training deep neural networks for video classification, the most commonly used approach to perform transfer learning is to apply finetuning from a similar dataset or problem. In cross-dataset transfer learning, the performed tasks are similar, but datasets show a difference in data distributions, label-space differences, or in the amount of available training data samples per class. It is reported that transferring knowledge from larger datasets for isolated SLR often improves accuracy [16]. However, when the datasets are similarly small in size and difference in content language, content or shots exists, little benefit is observed [27]. Although transfer learning datasets exist for video classification [28], there is no SL transfer dataset to test different transfer strategies.

In order to understand what information can be transferred in isolated SLR successfully with what method, a dataset that would allow us to experiment with closed and partial set transfer learning settings was necessary. Using such a dataset, it would become easier to independently test cross-dataset sign language transfer using several

state-of-the-art transfer learning approaches. Such datasets exist for similar domains such as image classification [145], video classification [28], but not sign language recognition. As such a dataset existed for video classification with the HMDB [146] and UTF-101 [147] datasets, we follow similar procedures to construct a baseline subset of the BSign22k and AUTSL datasets. These datasets share 57 common gestures. However, not all gestures are identical, as their performance varies from dataset to dataset due to differences in interpretation and style. Cross dataset sign language recognition can approximate and measure such differences in sign and signer characteristics. Among these two datasets, BSign22k contains fewer samples per class, so transfer learning from AUTSL to BSign22k is chosen as the use case that makes more sense.

Using this setting, we experiment with three different transfer learning settings. We construct a transfer learning pipeline using the SL-GCN algorithm as our baseline feature extractor. We use five different transfer learning techniques, namely, finetuning, DANN, MCC, JAN and DSBN. In setting 1, common signs between source and target datasets are unknown. In setting 2, a closed set transfer learning problem is observed where source and target class labels are identical and known. Setting 3 is a partial set transfer learning problem where the target set contains a subset of class labels belonging to the source set. In all three settings, we try to find transfer learning approaches that yield the largest improvement compared to baseline approaches.

7.1. Technical Approach

7.1.1. SL-GCN based Isolated Sign Language Recognition

In this section, we discuss the components of our isolated SLR pipeline. We perform coordinate-based Sign Language Recognition using coordinates extracted with the OpenPose library [65] from each dataset.

For isolated sign language recognition, we have a set of N training examples from each sign language dataset T as D_i^t, L_i^t, where D_i^t is a set of coordinates $D \in (J, T, 3)$ extracted from a single isolated SLR video clip, J is the number of joints, T is clip length, and the 3 values are the detected X, Y coordinates and the detection confidence values for each joint at each timestep. We obtain $J = 30$ joints belonging to fingertips, finger bases, wrists, arms, neck, mouth, nose, and eyes for each frame of a sign language video.

Having obtained the joints for each gesture, we apply several normalizations and augmentations to make our models more robust to small changes in user performance. A typical property of the isolated SLR datasets we use is that they both share the same rest pose where hands rest to the side of the users' legs. Frames that contain stationary hands in this rest pose are trimmed from the beginning and end of each video segment. In addition, further sampling is done from the sampled frames to bring the number of total frames sampled from each video to J frames. This was done using a random uniform sampling from a set of fixed intervals. This approach created duplicates of frames for shorter clips and provided temporal variation when sampling from longer clips.

For each coordinate, we apply a spatial normalization where the origin of the 2D coordinate system is moved to the temporal mean of the neck joint for that gesture. In addition, random horizontal mirroring and random spatial coordinate translation augmentations were found to improve recognition performance and were thus added to our pipeline.

The classifier base of our model is the SL-GCN model proposed by Jiang *et al.* [16]. The model takes input as a sequence of fixed length coordinates and outputs class prediction probabilities for sign language gestures.

We model the problem so that nodes of the graph correspond to joint landmark locations. The spatio-temporal graph adjacency matrix A is constructed in the spatial

domain according to anatomical spatial ordering, where neighboring joints are assigned a value of 1 and all other joints are assigned a value of 0. In the temporal domain, all the joints are only connected to themselves.

The architecture of the model consists of ten SL-GCN blocks for node and edge processing. Each SL-GCN block consists of a spatial convolution layer, multiplication with the adjacency matrix, and temporal convolution layer. The model includes decoupling with drop graph module proposed in [95] and spatio temporal channel (STC) attention modules proposed in [202]. The decoupling adds increased recognition power by dropping random joints along with their neighbors closer than an adjacency of K. Similarly, the STC attention module increases recognition power by focusing more on important joints, frames, and coordinates for certain gestures.

7.1.2. Supervised Transfer Methods

For each different transfer learning setup, this baseline feature extractor is used with different loss functions and classifiers in our experiments. We use several different transfer learning methods that aim to perform transfer learning by focusing on different characteristics of the data.

The DANN approach proposed in [134] promotes the use of a gradient reversal layer to learn domain-independent features. In transfer learning setting, the model takes half the samples in each mini-batch from the source and target domains. The model has two classifiers, one for class prediction and the other for dataset prediction. The model learns to make accurate label predictions from samples of both datasets while adversarially trying to suppress any features that can allow the model to distinguish between samples of data from different datasets, such as user physical or performance-related characteristics.

The DSBN approach shares all model parameters except batch normalization layers within the feature encoder. The batch normalization layer in a neural network

regularizes feature representations from different domains without taking into account class or domain information. However, when domain discrepancy is significant, the effect of batch normalization is diminished. In this approach, individual batch normalization layers keep track of unique normalization parameters and batch statistics values for each domain.

The JAN approach proposed in [132] maps features from different domains into a new data space where inter-class features have a more significant similarity. The method proposed, named joint maximum mean discrepancy (JMMD), minimizes the joint probability distribution distance of the source and target class-specific layers. The approach adds task-specific layers on top of the base SL-GCN network to learn mapping a common domain.

In a different approach, rather than using straightforward entropy minimization, the Minimum Class Confusion (MCC) approach proposed by [133] attempts to utilize the structure of the classification output matrix to perform transfer learning. The approach minimizes classification error by minimizing pairwise class confusion among unlabeled target data within a mini-batch. The model takes as input the logit matrix and multiplies the logit matrix by its transpose to approximate a confusion matrix where correlated classes yield higher pairwise scores for each other more often. Probability rescaling and uncertainty reweighing are used to emphasize samples that are more important for classification. In contrast, category normalization is used to balance the effect of each class in a mini-batch. Finally, a loss function is calculated that maximizes the diagonal of the correlation matrix and minimizes the confusion of each class.

7.2. BSgin57 Subset

In this study, we have created a selection of similar isolated sign language signs from two publicly available sign language datasets. Going over the signs based on meaning and visual similarity, we selected 57 different signs from among 216 AUTSL

and 744 Bsign22k gestures. For BosphorusSign22k, we also divided the dataset by number of users to create two distinct test settings with single ($BSign22k_{single}$) and multi-user($BSign22k_{shared}$) target training settings. The details of each dataset are given in Table 7.1.

Table 7.1. Number of identical sign pairs and total number of videos from the AUTSL and Bsign22k datasets

	# signs	# train videos	# val.videos	# train users	# val. users
Bsign22k	744	17090	5452	4	2
AUTSL	216	28139	3742	31	6
$BSign22k_{shared}$	57	1496	498	4	2
$BSign22k_{single}$	57	377	498	1	2
$AUTSL_{shared}$	57	7076	935	31	6

In the dataset selection process, 40 of the 57 selected signs are composed of signs that are performed in nearly perfect similarity. To increase domain difference between the two sets, the remaining 17 signs selected from the datasets show slight differences in hand orientation, movement direction, or the presence of additional morphemes as part of a compound gesture.

7.3. Experiments and Results

We evaluate our methods on the BSign22k and AUTSL datasets. Below we introduce and detail the transfer learning subsets derived from these datasets for each experiment. First, we evaluate the existing datasets and baseline recognition methods to choose hyperparameters and optimize our ISLR pipeline. Next, we evaluate the cross-domain adaptation experiments on the $BSign22k_{shared}$ $BSign22k_{single}$ and $AUTSL_{shared}$ datasets. Then we move on to transfer from the larger dataset, and we evaluate whether partial set transfer learning from the AUTSL dataset is beneficial.

7.3.0.1. Baseline Isolated Recognition Results. To thoroughly evaluate the isolated recognition performance, we first evaluate the baseline recognition performance of the SL-GCN-based recognition method by itself using several different single source and finetuning-based approaches.



Figure 7.1. Representative Images from the $\text{BSign22k}_{\text{shared}}$ (images to the left with green background) and $\text{AUTSL}_{\text{shared}}$ (images to the right with white & complex backgrounds) datasets.

In Table 7.2, single dataset and cross dataset accuracy results from each dataset are reported. Since we assume that no knowledge of shared classes is present, the final fully connected layer of the model, which acts as the classifier, is discarded when transferring the pretrained layer. Results show that when transferring from the larger AUTSL dataset to the minor BSign22k dataset, improvement is observed. Improvement is more significant on the $\text{BSign22k}_{\text{shared}}$ dataset, where the total number of training videos is even smaller. On the other hand, transfer from Bsign22k to AUTSL yields minimal improvement.

Table 7.2. Baseline accuracy results for isolated sign language recognition when no prior information on shared classes to be transferred is present.

	AUTSL	**BSign**	**AUTSL$_{\text{shared}}$**	**BSign$_{\text{shared}}$**
Target Only	91.22	89.25	97.78	92.97
Finetuning on AUTSL	-	89.86	-	95.78
Finetuning on BSIGN22K	91.38	-	97.96	-

7.3.0.2. Supervised Closed Set Transfer Learning on Bsign22k. In this experimental setting, both the source and target datasets contain identical labels from 57 classes. Knowing the shared classes between the two datasets enables us to use supervised transfer learning approaches to facilitate better knowledge transfer between the datasets. The **target only** setting explores the accuracy of the baseline method without any input from the larger source domain. Likewise, the **source only** setting reports the accuracy of the model trained on the source domain without seeing the target domain. The two initial baseline transfer methods are the **combined** and **finetuning** approaches. In **combined**, half of the samples in each mini-batch are sampled from respective source and target domains, and the model is trained with a single cross-entropy loss. In finetuning, differing from 7.3.0.1 where we discarded the classifier fully connected layer, all layers are utilized during transfer. During finetuning, we initially freeze the feature extractor for ten epochs before unfreezing all layers and resuming training.

In Table 7.3, we report on closed set transfer learning results for the transfer learning setting where AUTSL$_{\text{shared}}$ is set as the source task and BSign22k$_{\text{shared}}$ and BSign22k$_{\text{single}}$ subsets of the dataset are set as the target. When the number of examples in the target training set is small, closed set transfer learning is more advantageous (the single-user case). In such a setting, the JAN method surpasses the baseline finetuning accuracy with 88.48%. When the number of training samples increases by setting the

Table 7.3. Top-1 Accuracy results for closed set transfer learning are reported. Each method uses $\text{AUTSL}_{\text{shared}}$ as source and $\text{BSign22k}_{\text{shared}}$ as target domains.

	Train	**Validation**	**Train-Target**	
	Source	**Target**	**$\text{BSign22k}_{\text{single}}$**	**$\text{BSign22k}_{\text{shared}}$**
target only		$\text{BSign22k}_{\text{shared}}$	68.28	92.97
source only	$\text{AUTSL}_{\text{shared}}$		65.26	65.26
combined	$\text{AUTSL}_{\text{shared}}$	$\text{BSign22k}_{\text{shared}}$	85.11	95.8
finetuning	$\text{AUTSL}_{\text{shared}}$	$\text{BSign22k}_{\text{shared}}$	85.34	96.19
DANN	$\text{AUTSL}_{\text{shared}}$	$\text{BSign22k}_{\text{shared}}$	86.41	96.54
MCC	$\text{AUTSL}_{\text{shared}}$	$\text{BSign22k}_{\text{shared}}$	86.84	**97.15**
JAN	$\text{AUTSL}_{\text{shared}}$	$\text{BSign22k}_{\text{shared}}$	**88.48**	94.23
DSBN	$\text{AUTSL}_{\text{shared}}$	$\text{BSign22k}_{\text{shared}}$	84.34	96.82

target task as $\text{BSign22k}_{\text{shared}}$, the increase from baseline methods is smaller. In that setting, MCC beats baseline finetuning and combined approaches, achieving 97.15% accuracy, while methods such as JAN and DANN show a negligible increase.

7.3.0.3. Partial Set Transfer Learning on Bsign22k. Utilizing transfer from a dataset with a larger vocabulary is also a common use case in supervised transfer learning. The source domain contains 57 common signs, but there are 159 additional classes. Class labels for both datasets are aligned so that all methods in this setting use the same 159 class classifier. The results of these experiments are presented in Table 7.4. With all baseline and transfer learning approaches, partial-set transfer from the larger AUTSL training set surpasses the transfer results from the $\text{AUTSL}_{\text{shared}}$ subset. For the $\text{BSign22k}_{\text{single}}$ and $\text{BSign22k}_{\text{shared}}$ subsets, the accuracy figures reach 90.56% and 98.63 %, respectively with the MCC algorithm. These results show a five and one percent improvement over transfer learning baselines with combined and finetuning-based training approaches with the same experimental setting.

Table 7.4. Top-1 Accuracy results for partial-set transfer learning are reported. Each method uses AUTSL as source and $\text{BSign22k}_{\text{shared}}$ as target domains.

	Train	**Validation**	**Train-Target**	
	Source	**Target**	$\textbf{BSign22k}_{\textbf{single}}$	$\textbf{BSign22k}_{\textbf{shared}}$
target only		$\text{BSign22k}_{\text{shared}}$	68.28	92.97
source only	AUTSL		62.79	*71.23*
combined	AUTSL	$\text{BSign22k}_{\text{shared}}$	85.11	98.12
finetuning	AUTSL	$\text{BSign22k}_{\text{shared}}$	88.75	97.14
DANN	AUTSL	$\text{BSign22k}_{\text{shared}}$	88.75	98.19
MCC	AUTSL	$\text{BSign22k}_{\text{shared}}$	**90.56**	**98.63**
JAN	AUTSL	$\text{BSign22k}_{\text{shared}}$	90.16	96.72

In the baseline methods in Table 7.4, the source-only method achieves lower accuracy than the same experiment with the closed set transfer learning method. As the base source, only classifier in partial-set transfer learning has more classes, it can make more mistakes. In addition, as the size of the source dataset increases, the transfer approaches DANN, MCC and JAN yield scores that are closer to baseline approaches such as combined and finetuning approaches.

Table 7.5. Top-1 Accuracy results for fusion of transfer learning methods. Each method uses AUTSL as source and $\text{BSign22k}_{\text{shared}}$ as target domains.

	Training		**Validation**
	Source	**Target**	$\text{BSign22k}_{\text{shared}}$
finetuning + DANN	AUTSL	$\text{BSign22k}_{\text{shared}}$	97.14
finetuning + MCC	AUTSL	$\text{BSign22k}_{\text{shared}}$	**98.59**
finetuning + JAN	AUTSL	$\text{BSign22k}_{\text{shared}}$	96.91
finetuning + DSBN	AUTSL	$\text{BSign22k}_{\text{shared}}$	98.19
finetuning + DANN + MCC	AUTSL	$\text{BSign22k}_{\text{shared}}$	92.57
finetuning + JAN + MCC	AUTSL	$\text{BSign22k}_{\text{shared}}$	84.33
finetuning + DSBN + MCC	AUTSL	$\text{BSign22k}_{\text{shared}}$	**98.8**

In Table 7.5, we explore the combinations of these classifiers with finetuning and each other on the $\text{BSign22k}_{\text{shared}}$ dataset. The fusion of these algorithms is achieved by initializing the feature extractor and classifier layers of the algorithm, freezing them for the first five epochs, and then applying respective model architecture loss combinations to a single model. Of the attempted methods, finetuning and MCC approaches yield the most promising results. In a greedy fashion, we combined this method with several other methods, which gave us 98.8% accuracy on the $\text{BSign22k}_{\text{shared}}$ task of the dataset.

8. CONCLUSION

In this thesis, we addressed the topic of isolated Sign Language Recognition, focusing on the development of source-independent representation and recognition algorithms using transfer learning. To do this, we concentrated on three open research areas: constructing efficient baseline representations, conducting source-independent temporal alignments for isolated signs, and recognizing sign language across datasets.

In chapter 3, the thesis focused on building a baseline dataset for isolated Turkish Sign Language recognition BSign22k and obtaining baseline recognition results. 3D convolutional neural networks were determined to be the optimal solution for this task. Leave one user out experiments on the dataset revealed that discrepancies in user sign performances were a key reason for the suggested method's poor performance on some signs.

In order to remedy this, in chapter 4, we presented the ATAF method, which is a pose-based visual representation built on the idea of aggregating joint heatmaps over isolated sign language videos. The method focused on building signer and dataset-independent concise linguistic representations which relied on modeling signs' dynamic and static subunits. We proposed a keyframe-based linguistic temporal colorization scheme that represents static frames in hand and finger movement to describe compound signs as a collection of distinct subunits. To detect these keyframes, we implemented a static subunit detection method that uses local minima in hand speed with density clustering to extract dynamic and static subunits. We use an end-to-end alignment approach to align subunits so that extracted keyframes from differently performed signs can be aligned with respect to speed and joint coordinate locations. We use additional features and joints, including detailed finger joint coordinates, and introduce an end-to-end training scheme by optimizing the ATAF features for fast and efficient extraction. An ablation study on feature construction is performed to find the best-performing features and parameter optimizations for using different joint groups

and model architectures.

In terms of aligning keyframes, we make use of end-to-end trained Temporal Transformers to align keyframes and linearly sampled frames. Analysis of keyframe extraction schemes with different features demonstrates the effectiveness of the proposed approach. The experiments demonstrated that compound signs composed of two or more signs benefited from keyframe-based approaches in contrast to repetitive and circular signs that benefited more from longer sequences of aligned uniformly sampled signs. Our analysis demonstrated that modeling subunits with keyframes improved accuracy in sign types that adhered well to the movement-hold model, such as non-circular and compound gestures, showing larger-than-average accuracy. In circular and mono-morphemic gestures, the accuracy obtained with features constructed by linearly sampled and aligned frames was considerably higher.

Signs are composed of complicated manual and non-manual configurations with many degrees of freedom, where learning complicated representations from RGB images requires huge amounts of annotated data. Since the available sign language datasets were orders of magnitude smaller than the action recognition datasets both in variety and size, we demonstrated that it is possible to improve on the state-of-the-art in isolated SLR by combining state-of-the-art 3D-CNN methods with complementary features where joint coordinates of the person performing the sign language and the spatio-temporal subunits of the performed gestures are emphasized. On the public AUTSL dataset, the method achieves 92.35% recognition accuracy, beating competitive skeleton-based approaches. Similarly, the method achieves a peak accuracy of 80.14% over 744 classes of the BSign22k dataset, which is higher than the 71.02% performance achieved with linear sampling and 51.2% achieved with the comparable Potion method [14]. The proposed method, which only uses joint coordinates, aims to serve a complementary role to that of 3D convolutional SLR approaches. Score fusion of ATAF with MC3 ResNet models [74] yields the top performance of 94.9% accuracy on the BSign22k and 97.6% accuracy on the AUTSL, surpassing top accuracy results obtained by methods using only the RGB image as input.

In chapter 5, we establish a framework for transfer learning on temporal data and use a multi-modal prediction framework for personality prediction to develop methods for spatio-temporal transfer learning. We use those methods in chapter 6, to utilize the multi-task learning framework for sign language alignment and recognition. We explore the usage of Canonical Correlation based temporal alignment methods to perform temporal registration on signs. The project's primary objective was to investigate the use of temporal registration to enhance the recognition performance of traditional techniques such as RDFs and 3D convolutional neural networks. Experiments showed that temporally registering signs prior to categorization improves recognition accuracy in comparison to brute-force approaches such as temporal sliding windows. In addition, the task-aware CTW method we presented further increases performance by clustering the gestures in the dataset into subgroups independent of class or user information. By aligning similar gestures such as longer gestures, shorter gestures or gestures with a different number of movement and hold segments together, the proposed method found common groupings among the samples of the training set to make more natural alignments.

Finally, in chapter 7, we established a common sign language vocabulary subset from two publicly available Turkish Sign Language datasets and introduced experimental protocols for supervised transfer learning experiments. We believe the dataset will be a useful benchmark for testing novel supervised and unsupervised transfer learning methods for video classification. Using that dataset, we proposed a sign language classification methodology that uses graph convolutional neural networks and deep transfer learning mechanisms to improve isolated SLR performance on under-resourced sign language datasets. We experimented with three protocols: closed-set transfer learning, partial-set transfer learning, and transfer learning without knowledge of shared classes between tasks. Experimental results show that when shared class knowledge is present, supervised transfer learning techniques improve the performance of isolated SLR. The observed improvement is more significant in the single-user test. In the case of closed set transfer learning, the improvements in performance over baseline methods can be attributed to the increased number of samples per class. Improvements with

transfer methods such as MCC, DSBN, Adversarial Discriminative Domain Adaptation (ADDA), and JAN show that improved transfer efficiency from the same class samples of a different domain also leads to further improvement. In addition, in the case of partial-set transfer learning, the benefit observed from the addition of source data belonging to out of vocabulary gestures becomes apparent as the number of signs in the training set increases and class knowledge is preserved during transfer. In addition, observed improvements with the fusion of MCC, finetuning, and DSBN show that applying transfer methods that focus on different aspects of neural network models such as normalization layers, initialization, and loss functions further increase the benefit that can be observed from utilizing shared class knowledge.

Sign language recognition is a challenging research area where annotated data is an important ingredient for success. Future directions of research in sign language recognition will consequently focus on exploiting supervised and semi-supervised annotated data sources to develop improved representations and modeling strategies for signer and dataset-independent isolated and continuous recognition systems. Of the methods we proposed, ATAF and 3D Convolutional neural network based approaches can be extended to continuous data with sliding window approaches or by using extracted features with temporal classifiers. In addition, the proposed partial transfer learning approaches are suitable for continuous SLR data as well. Such methods can be used to extend our algorithm to other datasets, where data distributions might be different from the isolated SLR datasets we used in this thesis. Transfer learning methods can be used to extend our models to other domains such as video-chat where signers face restrictions such as signing all gestures one-handed because of a need to hold the phone in the other hand. Likewise, SLR in the wild or SLR for children are some interesting use cases with little available annotated data that can be good candidates for under-resourced SLR. In addition, in this dissertation we annotated and extracted phonological properties which were useful for analyzing the strengths and weaknesses of our methods. While using the four phonological properties we identified was not helpful in isolated recognition, other properties such as selected fingers and sign types can be inferred from body joints and may also be useful for recognition. As one possible

extension to our work, we plan on using multi-task and multi-label learning approaches to incorporate phonological properties into SLR classification pipelines.

REFERENCES

1. Easterbrooks, S., A. Lederberg, S. Antia, B. Schick, P. Kushalnagar, M.-Y. Webb, L. Branum-Martin and C. M. Connor, "Reading Among Diverse DHH Learners: What, How, and for Whom?", *American Annals of the Deaf*, Vol. 159, pp. 419–432, 2015.

2. Starner, T. and A. Pentland, "Real-Time American Sign Language Recognition from Video Using Hidden Markov Models", *Proceedings of International Symposium on Computer Vision*, pp. 265–270, IEEE, Florida, USA, 1995.

3. Kadir, T., R. Bowden, E.-J. Ong and A. Zisserman, "Minimal Training, Large Lexicon, Unconstrained Sign Language Recognition.", *British Machine Vision Conference*, pp. 1–10, Cardiff, UK, 2004.

4. Camgoz, N. C., O. Koller, S. Hadfield and R. Bowden, "Multi-channel Transformers for Multi-articulatory Sign Language Translation", *Lecture Notes in Computer Science*, Vol. 12538, pp. 301–319, 2020.

5. Wittenburg, P., H. Brugman, A. Russel, A. Klassmann and H. Sloetjes, "ELAN: A Professional Framework for Multimodality Research", *5th International Conference on Language Resources and Evaluation*, pp. 1556–1559, Lisbon, Portugal, 2006, `https://archive.mpi.nl/tla/elan`.

6. Jahn, E., R. Konrad, G. Langer, S. Wagner and T. Hanke, "Publishing DGS Corpus Data: Different Formats for Different Needs", *Proceedings of the Workshop on the Representation and Processing of Sign Languages at International Conference on Language Resources and Evaluation*, Vol. 2, Miyazaki, Japan, 2018.

7. Albanie, S., G. Varol, L. Momeni, H. Bull, T. Afouras, H. Chowdhury, N. Fox, B. Woll, R. Cooper, A. McParland *et al.*, "BBC-Oxford British Sign Language

Dataset", *arXiv preprint arXiv:2111.03635*, 2021.

8. Neidle, C., A. Thangali and S. Sclaroff, "Challenges in Development of the American Sign Language Lexicon Video Dataset (ASLLVD) Corpus", *International Conference on Language Resources and Evaluation 2012*, pp. 143–150, European Language Resources Association (ELRA), Istanbul, Turkey, 2012.

9. Moryossef, A., I. Tsochantaridis, J. Dinn, N. C. Camgoz, R. Bowden, T. Jiang, A. Rios, M. Muller and S. Ebling, "Evaluating the Immediate Applicability of Pose Estimation for Sign Language Recognition", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3434–3440, Nashville, USA, 2021.

10. Özdemir, O., A. A. Kındıroğlu, N. C. Camgöz and L. Akarun, "Bosphorussign22k Sign Language Recognition Dataset", *arXiv preprint arXiv:2004.01283*, 2020.

11. Camgoz, N. C., A. A. Kindiroglu and L. Akarun, "Gesture Recognition Using Template Based Random Forest Classifiers", *European Conference on Computer Vision 2014 Workshops, Pt I*, 2014.

12. Neverova, N., C. Wolf, G. Taylor and F. Nebout, "ModDrop: Adaptive Multi-Modal Gesture Recognition", *Transactions on Pattern Analysis and Machine Intelligence*, Vol. 38, pp. 1692–1706, 2016.

13. Carreira, J. and A. Zisserman, "I3D for Kinetics", *30th IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 2017-January, pp. 4724–4733, 2017.

14. Choutas, V., P. Weinzaepfel, J. Revaud and C. Schmid, "PoTion: Pose MoTion Representation for Action Recognition", *Proceedings of the Conference on Computer Vision and Pattern Recognitio*, pp. 7024–7033, Salt Lake City, USA, 2018.

15. Koller, O., H. Ney and R. Bowden, "Deep Hand: How to Train a CNN on 1 Million

Hand Images When Your Data is Continuous and Weakly Labelled", *2016 IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, 2016.

16. Jiang, S., B. Sun, L. Wang, Y. Bai, K. Li and Y. Fu, "Skeleton Aware Multi-Modal Sign Language Recognition", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3413–3423, Nashville, USA, 2021.

17. Camgöz, N. C., O. Koller, S. Hadfield and R. Bowden, "Sign Language Transformers: Joint End-to-end Sign Language Recognition and Translation", *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 10020–10030, IEEE Computer Society, Seattle, USA, 2020.

18. Tibshirani, R., "Regression Shrinkage and Selection via the Lasso", *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 58, No. 1, pp. 267–288, 1996.

19. Zhou, J., J. Chen and J. Ye, "MALSAR: Multi-task Learning via Structural Regularization", Arizona State University, Tempe, Arizona, 2011, `http://http://jiayuzhou.github.io/MALSAR/`, accessed on August 10, 2022.

20. Jolliffe, I. T., *Principal Component Analysis for Special Types of Data*, Springer, New York, 2002.

21. Peng, H., F. Long and C. Ding, "Feature Selection Based on Mutual Information Criteria Of Max-Dependency, Max-Relevance, and Min-Redundancy", *Transactions on Pattern Analysis and Machine Intelligence*, Vol. 27, No. 8, pp. 1226–1238, 2005.

22. Zhou, F. and F. Torre, "Canonical Time Warping for Alignment of Human Behavior", *Advances in Neural Information Processing Systems*, Vol. 22, pp. 2286–2294, Vancouver, Canada, 2009.

23. Zhou, F. and F. De la Torre, "Generalized Time Warping for Multi-Modal Alignment of Human Motion", *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1282–1289, Providence, USA, 2012.

24. Sincan, O. M. and H. Y. Keles, "Using Motion History Images with 3D Convolutional Networks in Isolated Sign Language Recognition", *arXiv preprint arXiv:2110.12396*, 2021.

25. Kay, W., J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "Kinetics Human Action Video Dataset", *arXiv preprint arXiv:1705.06950*, 2017.

26. Monfort, M., A. Andonian, B. Zhou, K. Ramakrishnan, S. A. Bargal, T. Yan, L. Brown, Q. Fan, D. Gutfreund, C. Vondrick *et al.*, "Moments in Time Dataset: One Million Videos for Event Understanding", *Transactions on Pattern Analysis and Machine Intelligence*, Vol. 42, No. 2, pp. 502–508, 2019.

27. Selvaraj, P., G. NC, P. Kumar and M. Khapra, "OpenHands: Making Sign Language Recognition Accessible with Pose-based Pretrained Models across Languages", *arXiv preprint arXiv:2110.05877.*, 2021.

28. Chen, M.-H., Z. Kira, G. AlRegib, J. Yoo, R. Chen and J. Zheng, "Temporal Attentive Alignment for Large-Scale Video Domain Adaptation", *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6321–6330, Long Beach, USA, 2019.

29. Camgöz, N. C., A. A. Kındıroğlu, S. Karabüklü, M. Kelepir, L. Akarun and S. Ozsoy, "BosphorusSign : A Turkish Sign Language Recognition Corpus in Health and Finance Domains", *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pp. 1383–1388, Portoroz, Slovenia, 2016.

30. Özdemir, O., A. A. Kındıroğlu, N. C. Camgöz and L. Akarun, "BosphorusSign22k

Sign Language Recognition Dataset", *Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages. International Conference on Language Resources and Evaluation*, pp. 181–188, Marseille, France, 2020.

31. Kindiroglu, A. A., O. Ozdemir and L. Akarun, "Temporal Accumulative Features for Sign Language Recognition", *Proceedings - 2019 International Conference on Computer Vision Workshop*, pp. 1288–1297, Institute of Electrical and Electronics Engineers Inc., Seoul, Korea, 2019.

32. Kindiroglu, A. A., O. Ozdemir and L. Akarun, "Aligning Accumulative Representations for Sign Language Recognition", *Machine Vision and Applications*, Vol. 34, No. 1, p. 12, 2023.

33. Stokoe, W. C., "Sign Language Structure", *Annual Review of Anthropology*, Vol. 9, No. 1, pp. 365–390, 1980.

34. Gökgöz, K., "Syllables in TİD", *Dilbilim Araştırmaları Dergisi*, Vol. 29, No. 1, pp. 29–49, 2018.

35. Battison, R., "Lexical Borrowing in American Sign Language", *Linguistics of American Sign Language: An Introduction*, p. 240, 1978.

36. Liddell, S. K. and R. E. Johnson, "American Sign Language: the Phonological Base", *Sign Language Studies*, Vol. 64, No. 1, pp. 195–277, 1989.

37. Cooper, H., B. Holt and R. Bowden, "Sign Language Recognition", *Visual Analysis of Humans*, Vol. 27, pp. 1–23, 2011.

38. Grobel, K. and M. Assan, "Isolated Sign Language Recognition Using Hidden Markov Models", *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, Vol. 1, pp. 162–167, IEEE, Orlando, FL, USA, 1997.

39. Tamura, S. and S. Kawasaki, "Recognition of Sign Language Motion Images",

Pattern Recognition, Vol. 21, No. 4, pp. 343–353, 1988.

40. Cui, Y. and J. Weng, "Appearance-Based Hand Sign Recognition from Intensity Image Sequences", *Computer Vision and Image Understanding*, Vol. 78, No. 2, pp. 157–176, 2000.

41. Camgoz, N. C., S. Hadfield, O. Koller, H. Ney and R. Bowden, "Neural Sign Language Translation", *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 7784–7793, IEEE Computer Society, Salt Lake City, USA, 2018.

42. Krizhevsky, A., I. Sutskever and G. E. Hinton, "Imagenet classification with Deep Convolutional Neural Networks", *Conference on Neural Information Processing Systems*, pp. 1097–1105, Lake Tahoe, USA, 2012.

43. He, K., X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition", *Proceedings of the Conference on Computer Vision and Pattern Recognition, 2016*, pp. 770–778, Las Vegas, USA, 2016.

44. Szegedy, C., P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich, "Going Deeper with Convolutions", *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 1–9, Boston, USA, 2015.

45. Simonyan, K. and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition", *International Conference on Learning Representations*, San Diego, USA, 2015.

46. Efthimiou, E. and S. E. Fotinea, "GSLC: Creation and Annotation of a Greek Sign Language Corpus for HCI", *Lecture Notes in Computer Science*, Vol. 4554, No. Part 1, pp. 657–666, 2007.

47. Athitsos, V., C. Neidle, S. Sclaroff, J. Nash, A. Stefan, Q. Yuan and A. Thangali, "The American Sign Language Lexicon Video Dataset", *2008 IEEE Computer*

Society Conference on Computer Vision and Pattern Recognition Workshops, Anchorage, Alaska, 2008.

48. Ong, E. J., H. Cooper, N. Pugeault and R. Bowden, "Sign Language Recognition using Sequential Pattern Trees", *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2200–2207, Providence, USA, 2012.

49. Escalera, S., X. Baró, J. Gonzàlez, M. a. Bautista, M. Madadi, M. Reyes, V. Ponce, H. J. Escalante, J. Shotton and I. Guyon, "ChaLearn Looking at People Challenge 2014: Dataset and Results", *European Conference on Computer Vision Workshop*, Zurich, Switzerland, 2014.

50. Chai, X., G. Li, Y. Lin, Z. Xu, Y. Tang, X. Chen and M. Zhou, "Sign language recognition and translation with kinect", *International Conference and Workshops on Automatic Face and Gesture Recognition*, Vol. 655, p. 4, Shanghai, China, 2013.

51. Kapuscinski, T., M. Oszust, M. Wysocki and D. Warchol, "Recognition of Hand Gestures Observed by Depth Cameras", *International Journal of Advanced Robotic Systems*, Vol. 12, p. 36, 2015.

52. Ronchetti, F., F. Quiroga and L. Lanzarini, "LSA64 : An Argentinian Sign Language Dataset", *Congreso Argentino de Ciencias de la Computacion*, San Luis, USA, 2016.

53. Gutierrez-Sigut, E., B. Costello, C. Baus and M. Carreiras, "LSE-SIGN: A Lexical Database for Spanish Sign Language", *Behavior Research Methods*, Vol. 48, No. 1, pp. 123–137, 2016.

54. Huang, J., W. Zhou, H. Li and W. Li, "Attention-Based 3D-CNNs for Large-Vocabulary Sign Language Recognition", *IEEE Transactions on Circuits and*

Systems for Video Technology, Vol. 29, No. 9, pp. 2822–2832, 2019.

55. Joze, H. R. V. and O. Koller, "MS-ASL: A Large-Scale Data Set and Benchmark for Understanding American Sign Language", *British Machine Vision Conference*, pp. 1–10, Kingston, UK, 2019.

56. Ebling, S., N. C. Camgöz, P. B. Braem, K. Tissi, S. Sidler-Miserez, S. Stoll, S. Hadfield, T. Haug, R. Bowden, S. Tornay, M. Razavi and M. Magimai-Doss, "Smile Swiss German Sign Language Dataset", *11th International Conference on Language Resources and Evaluation*, pp. 4221–4229, Miyazaki, Japan, 2019.

57. Li, D., C. R. Opazo, X. Yu and H. Li, "Word-Level Deep Sign Language Recognition from Video: A New Large-Scale Dataset and Methods Comparison", *IEEE Winter Conference on Applications of Computer Vision*, pp. 1448–1458, Institute of Electrical and Electronics Engineers Inc., Hawaii, USA, 2019.

58. Sincan, O. M. and H. Y. Keles, "AUTSL: A Large Scale Multi-Modal Turkish sign Language Dataset and Baseline Methods", *IEEE Access*, Vol. 8, pp. 181340–181355, 2020.

59. Sridhar, A., R. G. Ganesan, P. Kumar and M. Khapra, "Include: A Large Scale Dataset for Indian Sign Language Recognition", *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1366–1375, Lisbon, Portugal, 2020.

60. Albanie, S., G. Varol, L. Momeni, H. Bull, T. Afouras, H. Chowdhury, N. Fox, B. Woll, R. Cooper, A. McParland *et al.*, "BBC-Oxford British Sign Language Dataset", *arXiv preprint arXiv:2111.03635*, 2021.

61. Hernandez-Rebollar, J. L., R. W. Lindeman and N. Kyriakopoulos, "A Multi-Class Pattern Recognition System for Practical Finger Spelling Translation", *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces*, pp. 185–190, IEEE, Pittsburgh, USA, 2002.

62. Aran, O., *Vision Based Sign Language Recognition: Modeling and Recognizing Isolated Signs with Manual and Non-Manual Components*, Ph.D. Thesis, Bogazici University, 2008.

63. Hrúz, M., P. Campr, E. Dikici, A. A. Kındıroğlu, Z. Krňoul, A. Ronzhin, H. Sak, D. Schorno, H. Yalçın, L. Akarun *et al.*, "Automatic Fingersign-to-Speech Translation System", *Journal on Multimodal User Interfaces*, Vol. 4, No. 2, pp. 61–79, 2011.

64. Shotton, J., A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman and A. Blake, "Real-Time Human Pose Recognition in Parts from Single Depth Images", *Proceedings of the Conference on Computer Vision and Pattern Recognition*, Vol. 2, Colorado Springs, USA, 2011.

65. Cao, Z., T. Simon, S.-E. Wei and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7291–7299, San Juan, Puerto Rico, 2017.

66. Grishchenko, I. and R. Valentin Bazarevsky, "MediaPipe Holistic—Simultaneous Face, Hand and Pose Prediction, on Device", *AI GoogleBlog*, Vol. 15, p. 0, 2022.

67. Wang, H., A. Kläser, C. Schmid and C.-L. Liu, "Dense Trajectories and Motion Boundary Descriptors for Action Recognition", *International Journal of Computer Vision*, Vol. 103, pp. 60–79, 2013.

68. Karpathy, A., G. Toderici, S. Shetty, T. Leung, R. Sukthankar and L. Fei-Fei, "Large-Scale Video Classification with Convolutional Neural Networks", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1725–1732, Colombus, USA, 2014.

69. Simonyan, K. and A. Zisserman, "Two-stream Convolutional Networks for Action

Recognition in Videos", *Conference on Neural Information Processing Systems*, Vol. 1, pp. 568–576, MIT Press, Montreal, Canada, 2014.

70. Yue-Hei Ng, J., M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga and G. Toderici, "Beyond Short Snippets: Deep Networks for Video Classification", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4694–4702, Boston, USA, 2015.

71. Varol, G., I. Laptev and C. Schmid, "Long-Term Temporal Convolutions for Action Recognition", *Transactions on Pattern Analysis and Machine Intelligence*, Vol. 40, No. 6, pp. 1510–1517, 2017.

72. Tran, D., L. Bourdev, R. Fergus, L. Torresani and M. Paluri, "Learning Spatiotemporal Features with 3D Convolutional Networks", *International Conference on Computer Vision, 2015*, pp. 4489–4497, IEEE, Santiago, Chile, 2015.

73. Carreira, J. and A. Zisserman, "Quo Vadis, Action Recognition? a New Model and the Kinetics Dataset", *Proceedings of the Conference on Computer Vision and Pattern Recognition, 2017*, pp. 4724–4733, IEEE, San Juan, Puerto Rico, 2017.

74. Tran, D., H. Wang, L. Torresani, J. Ray, Y. LeCun and M. Paluri, "a Closer Look at Spatiotemporal Convolutions for Action Recognition", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6450–6459, Salt Lake City, USA, 2018.

75. Feichtenhofer, C., H. Fan, J. Malik and K. He, "Slowfast Networks for Video Recognition", *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6202–6211, Seoul, Korea, 2019.

76. Lin, J., C. Gan and S. Han, "TSM: Temporal Shift Module for Efficient Video Understanding", *Proceedings of the IEEE/CVF International Conference on Com-*

puter Vision, pp. 7083–7093, Seoul, Korea, 2019.

77. Tran, D., H. Wang, L. Torresani and M. Feiszli, "Video Classification with Channel-Separated Convolutional Networks", *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5552–5561, Seoul, Korea, 2019.

78. Tan, M. and Q. Le, "Efficientnet: Rethinking Model Scaling for Convolutional Neural Networks", *International Conference on Machine Learning*, pp. 6105–6114, Proceedings of Machine Learning Research, Long Beach, USA, 2019.

79. Feichtenhofer, C., "X3D: Expanding Architectures for Efficient Video Recognition", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 203–213, Seattle, USA, 2020.

80. Zhang, K., W.-L. Chao, F. Sha and K. Grauman, "Video Summarization with Long Short-Term Memory", *European Conference on Computer Vision*, Springer, Amsterdam, Netherlands, 2016.

81. Ji, Z., K. Xiong, Y. Pang and X. Li, "Video Summarization with Attention-Based Encoder–Decoder Networks", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 30, No. 6, pp. 1709–1717, 2019.

82. Li, Z., K. Gavrilyuk, E. Gavves, M. Jain and C. G. M. Snoek, "Video LSTM Convolves, Attends and Flows for Action Recognition", *Computer Vision and Image Understanding*, Vol. 166, pp. 41–50, 2018.

83. Donahue, J., L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko and T. Darrell, "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description", *Transactions on Pattern Analysis and Machine Intelligence*, Vol. 39, No. 4, pp. 677–691, 2017.

84. Liu, Z., J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin and H. Hu, "Video Swin Transformer", *Proceedings of the IEEE/CVF Conference on Computer Vision*

and Pattern Recognition, pp. 3202–3211, New Orleans, USA, 2022.

85. Wu, Z., Y.-G. Jiang, X. Wang, H. Ye and X. Xue, "Multi-Stream Multi-Class Fusion of Deep Networks for Video Classification", *Proceedings of the 2016 ACM on Multimedia Conference*, pp. 791–800, ACM, Amsterdam, Netherlands, 2016.

86. Wang, L., Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang and L. Van Gool, "Temporal Segment Networks for Action Recognition in Videos", *Transactions on Pattern Analysis and Machine Intelligence*, Vol. 41, No. 11, pp. 2740–2755, 2018.

87. Albanie, S., G. Varol, L. Momeni, T. Afouras, J. S. Chung, N. Fox and A. Zisserman, "BSL-1K: Scaling Up Co-articulated Sign Language Recognition Using Mouthing Cues", *European Conference on Computer Vision*, Vol. 12356, pp. 35–53, Glasgow, Scotland, 2020.

88. Koller, O., N. C. Camgoz, H. Ney and R. Bowden, "Weakly Supervised Learning with Multi-Stream CNN-LSTM-HMMs to Discover Sequential Parallelism in Sign Language Videos", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 42, No. 9, pp. 2306–2320, 2019.

89. Contributors, M., "OpenMMLab Pose Estimation Toolbox and Benchmark", `https://github.com/open-mmlab/mmpose`, 2020.

90. Zhu, W., C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen and X. Xie, "Co-Occurrence Feature Learning for Skeleton Based Action Recognition Using Regularized Deep LSTM networks", *arXiv preprint arXiv:1603.07772*, 2016.

91. Mercanoglu Sincan, O. and H. Y. Keles, "Using Motion History Images with 3D Convolutional Networks in Isolated Sign Language Recognition", *IEEE Access*, Vol. 10, pp. 18608–18618, 2022.

92. Boháček, M. and M. Hrúz, "Sign Pose-Based Transformer for Word-Level Sign Language Recognition", *Proceedings of the IEEE/CVF Winter Conference on*

Applications of Computer Vision, pp. 182–191, Hawaii, USA, 2022.

93. Selvaraj, P., G. NC, P. Kumar and M. Khapra, "OpenHands: Making Sign Language Recognition Accessible with Pose-Based Pretrained models Across Languages", *arXiv preprint arXiv:2110.05877*, 2021.

94. Yan, S., Y. Xiong and D. Lin, "Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition", *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32, Louisiana, USA, 2018.

95. Cheng, K., Y. Zhang, C. Cao, L. Shi, J. Cheng and H. Lu, "Decoupling Gcn with Dropgraph Module for Skeleton-Based Action Recognition", *European Conference on Computer Vision*, pp. 536–553, Springer, Glassgow, UK, 2020.

96. Asghari-Esfeden, S., M. Sznaier and O. Camps, "Dynamic Motion Representation for Human Action Recognition", *The IEEE Winter Conference on Applications of Computer Vision*, pp. 557–566, Snowmass Village, USA, 2020.

97. Gao, Z., G. Lu, C. Lyu and P. Yan, "Key-frame Selection for Automatic Summarization of Surveillance Videos: a Method of Multiple Change-Point Detection", *Machine Vision and Applications*, Vol. 29, No. 7, pp. 1101–1117, 2018.

98. Xiong, W., C.-M. Lee and R.-H. Ma, "Automatic Video Data Structuring Through Shot Partitioning and Key-Frame Computing", *Machine Vision and Applications*, Vol. 10, No. 2, pp. 51–65, 1997.

99. Fanfani, M., F. Bellavia and C. Colombo, "Accurate Keyframe Selection and Keypoint Tracking for Robust Visual Odometry", *Machine Vision and Applications*, Vol. 27, No. 6, pp. 833–844, 2016.

100. Mo, H., F. Yamagishi, I. Ide, S. Satoh and M. Sakauchi, "Key Shot Extraction and Indexing in A News Video Archive", *Institute of Electronics, Information and Communication Engineers Technical Report*, Vol. 105, No. 118, pp. 55–59, 2005.

101. Xu, W., Z. Miao, J. Yu and Q. Ji, "Action Recognition and Localization with Spatial and Temporal Contexts", *Neurocomputing*, Vol. 333, pp. 351–363, 2019.

102. Yang, R. and S. Sarkar, "Detecting Coarticulation in Sign Language Using Conditional Random Fields", *18th International Conference on Pattern Recognition*, Vol. 2, pp. 108–112, IEEE, Hong Kong, China, 2006.

103. Zhao, Z. and A. M. Elgammal, "Information Theoretic Key Frame Selection for Action Recognition.", *Proceedings of the British Machine Vision Conference*, pp. 1–10, Newcastle, UK, 2008.

104. Carlsson, S. and J. Sullivan, "Action Recognition by Shape Matching to Key Frames", *Workshop on Models Versus Exemplars in Computer Vision*, Vol. 1, p. 18, Citeseer, San Juan, Puerto Rico, 2001.

105. Lu, G., Y. Zhou, X. Li and P. Yan, "Unsupervised, Efficient and Scalable Key-Frame Selection for Automatic Summarization of Surveillance Videos", *Multimedia Tools and Applications*, Vol. 76, No. 5, pp. 6309–6331, 2017.

106. Tang, H., H. Liu, W. Xiao and N. Sebe, "Fast and Robust Dynamic Hand Gesture Recognition via Key Frames Extraction and Feature Fusion", *Neurocomputing*, Vol. 331, pp. 424–433, 2019.

107. Rodriguez, A. and A. Laio, "Clustering by Fast Search and Find Of Density Peaks", *Science*, Vol. 344, No. 6191, pp. 1492–1496, 2014.

108. Fang, G., X. Gao, W. Gao and Y. Chen, "A Novel Approach to Automatically Extracting Basic Units from Chinese Sign Language", *International Conference on Pattern Recognition, 2004*, Vol. 4, pp. 454–457, IEEE, Cambridge, UK, 2004.

109. Kong, W. and S. Ranganath, "Sign Language Phoneme Transcription with Rule-Based Hand Trajectory Segmentation", *Journal of Signal Processing Systems*, Vol. 59, No. 2, pp. 211–222, 2010.

110. Theodorakis, S., V. Pitsikalis and P. Maragos, "Dynamic–Static Unsupervised Sequentiality, Statistical Subunits and Lexicon for Sign Language Recognition", *Image and Vision Computing*, Vol. 32, No. 8, pp. 533–549, 2014.

111. Elakkiya, R. and K. Selvamani, "Extricating Manual and Non-Manual Features for Subunit Level Medical Sign Modelling in Automatic Sign Language Classification and Recognition", *Journal of Medical Systems*, Vol. 41, No. 11, pp. 1–13, 2017.

112. De Coster, M., M. Van Herreweghe and J. Dambre, "Sign Language Recognition with Transformer Networks", *12th International Conference on Language Resources and Evaluation*, pp. 6018–6024, European Language Resources Association (ELRA), Marseille, France, 2020.

113. Tavella, F., V. Schlegel, M. Romeo, A. Galata and A. Cangelosi, "WLASL-LEX: A Dataset for Recognising Phonological Properties in American Sign Language", *60th Annual Meeting of the Association for Computational Linguistics*, Vol. 2, pp. 453–463, Dublin, Ireland, 2022.

114. Berndt, D. J. and J. Clifford, "Using Dynamic Time Warping to Find Patterns in Time Series", *Proceedings of AAAI Workshop on Knowledge Discovery in Databases*, Vol. 10, pp. 359–370, Seattle, USA, 1994.

115. Cuturi, M. and M. Blondel, "Soft-DTW: A Differentiable Loss Function for Time-Series", *arXiv preprint arXiv:1703.01541*, 2017.

116. Petitjean, F., A. Ketterlin and P. Gançarski, "A Global Averaging Method for Dynamic Time Warping, with Applications to Clustering", *Pattern Recognition*, Vol. 44, No. 3, pp. 678–693, 2011.

117. Trigeorgis, G., M. A. Nicolaou, S. Zafeiriou and B. W. Schuller, "Deep Canonical Time Warping", *Proceedings of the IEEE Conference on Computer Vision and*

Pattern Recognition, pp. 5110–5118, Las Vegas, USA, 2016.

118. Chang, C.-Y., D.-A. Huang, Y. Sui, L. Fei-Fei and J. C. Niebles, "D3TW: Discriminative Differentiable Dynamic Time Warping for Weakly Supervised Action Alignment and Segmentation", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3546–3555, Long Beach, USA, 2019.

119. Korbar, B., D. Tran and L. Torresani, "Scsampler: Sampling Salient Clips from Video for Efficient Action Recognition", *Proceedings of the IEEE International Conference on Computer Vision*, pp. 6232–6242, Seoul, Korea, 2019.

120. Lohit, S., Q. Wang and P. Turaga, "Temporal Transformer Networks: Joint Learning of Invariant and Discriminative Time Warping", *Computer Vision and Pattern Recognition*, p. 12418, IEEE/CVF, Long Beach, USA, 2019.

121. Oh, J., J. Wang and J. Wiens, "Learning to Exploit Invariances in Clinical Time-Series Data Using Sequence Transformer Networks", *arXiv preprint arXiv:1808.06725*, 2018.

122. Wang, M. and W. Deng, "Deep Visual Domain Adaptation: A Survey", *Neurocomputing*, Vol. 312, pp. 135–153, 2018.

123. Zhang, J., W. Li, P. Ogunbona and D. Xu, "Recent Advances in Transfer Learning for Cross-Dataset Visual Recognition: a Problem-Oriented Perspective", *ACM Computing Surveys*, Vol. 52, No. 1, pp. 1–38, 2019.

124. Ouali, Y., C. Hudelot and M. Tami, "An Overview of Deep Semi-Supervised Learning", *arXiv preprint arXiv:2006.05278*, 2020.

125. Wilson, G. and D. J. Cook, "A Survey of Unsupervised Deep Domain Adaptation", *ACM Transactions on Intelligent Systems and Technology (TIST)*, Vol. 11, No. 5, pp. 1–46, 2020.

126. Dai, W., Q. Yang and y. Yu, "Boosting for Transfer Learning", *Proceedings of International Conference on Machine Learning*, ICML, Oregon, USA, 2007.

127. Daumé III, H., "Frustratingly Easy Domain Adaptation", *Proceedings of the AAAI Conference on Artificial Intelligence*, 1, pp. 256–263, Stanford, USA, 2007.

128. Pan, S. J., I. W. Tsang, J. T. Kwok and Q. Yang, "Domain Adaptation via Transfer Component Analysis", *IEEE Transactions on Neural Networks*, Vol. 22, No. 2, pp. 199–210, 2010.

129. Yosinski, J., J. Clune, Y. Bengio and H. Lipson, "How Transferable Are Features in Deep Neural Networks?", *Advances in Neural Information Processing Systems*, Vol. 27, 2014.

130. Yun, S., D. Han, S. J. Oh, S. Chun, J. Choe and Y. Yoo, "Cutmix: Regularization Strategy To Train Strong Classifiers with Localizable features", *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6023–6032, Seoul, Korea, 2019.

131. Tzeng, E., J. Hoffman, N. Zhang, K. Saenko and T. Darrell, "Deep Domain Confusion: Maximizing for Domain Invariance", *arXiv preprint arXiv:1412.3474*, 2014.

132. Long, M., H. Zhu, J. Wang and M. I. Jordan, "Deep Transfer Learning with Joint Adaptation Networks", *International Conference on Machine Learning*, pp. 2208–2217, Proceedings of Machine Learning Research, Sydney, Australia, 2017.

133. Jin, Y., X. Wang, M. Long and J. Wang, "Minimum Class Confusion for Versatile Domain Adaptation", *European Conference on Computer Vision*, pp. 464–480, Springer, Glassgow, UK, 2020.

134. Ganin, Y., E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand and V. Lempitsky, "Domain-Adversarial Training of Neural Net-

works", *The Journal of Machine Learning Research*, Vol. 17, No. 1, pp. 2096–2030, 2016.

135. Evgeniou, T. and M. Pontil, "Regularized Multi-Task Learning", *International Conference on Knowledge Discovery and Data Mining*, p. 109, ACM Press, New York, USA, 2004.

136. Argyriou, A., T. Evgeniou and M. Pontil, "Multi-Task Feature Learning", B. Schölkopf, J. Platt and T. Hoffman (Editors), *Advances in Neural Information Processing Systems*, Vol. 19, p. 1, MIT Press, Vancouver, Canada, 2006.

137. Gong, P., J. Ye and C. Zhang, "Robust Multi-Task Feature Learning.", *KDD : proceedings / International Conference on Knowledge Discovery & Data Mining. International Conference on Knowledge Discovery & Data Mining*, Vol. 2012, pp. 895–903, 2012.

138. Jalali, A., P. Ravikumar, S. Sanghavi and C. Ruan, "A Dirty Model for Multi-task Learning.", *Conference on Neural Information Processing Systems*, pp. 1–9, Vancouver, Canada, 2010.

139. Kang, Z., K. Grauman and F. Sha, "Learning with whom to share in multi-task feature learning", *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pp. 521–528, Washington, USA, 2011.

140. Chen, X., S. Kim, Q. Lin, J. G. Carbonell and E. P. Xing, "Graph-Structured Multi-Task Regression and An Efficient Optimization Method for General Fused Lasso", *arXiv preprint arXiv:1005.3579*, 2010.

141. Xu, Y., J. Yang, H. Cao, K. Mao, J. Yin and S. See, "Aligning Correlation Information for Domain Adaptation in Action Recognition", *arXiv preprint arXiv:2107.04932*, 2021.

142. Bellitto, G., F. Proietto Salanitri, S. Palazzo, F. Rundo, D. Giordano and

C. Spampinato, "Hierarchical Domain-Adapted Feature Learning for Video Saliency Prediction", *International Journal of Computer Vision*, Vol. 129, No. 12, pp. 3216–3232, 2021.

143. Sarhan, N. and S. Frintrop, "Transfer Learning for Videos: From Action Recognition to Sign Language Recognition", *Proceedings - International Conference on Image Processing, ICIP*, Vol. 2020-Octob, pp. 1811–1815, IEEE Computer Society, Abu Dhabi, United Arab Emirates, 2020.

144. Li, D., X. Yu, C. Xu, L. Petersson and H. Li, "Transferring Cross-Domain Knowledge for Video Sign Language Recognition", *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 6204–6213, Seattle, USA, 2020.

145. Ringwald, T. and R. Stiefelhagen, "Adaptiope: A Modern Benchmark for Unsupervised Domain Adaptation", *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 101–110, Snowmass Village, USA, 2021.

146. Kuehne, H., H. Jhuang, E. Garrote, T. Poggio and T. Serre, "HMDB: A Large Video Database for Human Motion Recognition", *2011 International Conference on Computer Vision*, pp. 2556–2563, IEEE, Barcelona, Spain, 2011.

147. Soomro, K., A. R. Zamir and M. Shah, "UCF101: A Dataset of 101 Human Actions Classes from Videos in the Wild", *arXiv preprint arXiv:1212.0402*, 2012.

148. Vazquez-Enriquez, M., J. L. Alba-Castro, L. Docio-Fernandez and E. Rodriguez-Banga, "Isolated Sign Language Recognition with Multi-Scale Spatial-Temporal Graph Convolutional Networks", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 3462–3471, Nashville, USA, 2021.

149. Zhang, Z., "Microsoft Kinect Sensor and Its Effect", *IEEE MultiMedia*, Vol. 19, No. 2, pp. 4–10, 2012.

150. Cao, Z., G. Hidalgo, T. Simon, S.-E. Wei and Y. Sheikh, "OpenPose: Realtime multi-Person 2D Pose Estimation Using Part Affinity Fields", *arXiv preprint arXiv:1812.08008*, 2018.

151. Berent, I., A. Dupuis and D. Brentari, "Amodal Aspects of Linguistic Design", *PLoS ONE*, Vol. 8, p. e60617, 2013.

152. Paszke, A., S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai and S. Chintala, "PyTorch: An Imperative Style, High-Performance Deep Learning Library", H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox and R. Garnett (Editors), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035, Curran Associates, Inc., Montreal, 2019.

153. Kingma, D. P. and J. Ba, "Adam: A Method for Stochastic Optimization.", *Proceedings of 3rd International Conference on Learning Representations*, San Diego, USA, 2014.

154. Pitsikalis, V., S. Theodorakis, C. Vogler and P. Maragos, "Advances in Phonetics-Based Sub-Unit Modeling for Transcription Alignment and Sign Language Recognition", *Conference on Computer Vision and Pattern Recognition 2011 Workshops*, pp. 1–6, Colorado Springs, USA, 2011.

155. Cooper, H., E.-J. Ong, N. Pugeault and R. Bowden, "Sign Language Recognition Using Sub-Units", *The Journal of Machine Learning Research*, Vol. 13, No. 1, pp. 2205–2231, 2012.

156. Alp Kindiroglu, A., O. Ozdemir and L. Akarun, "Temporal Accumulative Fea-

tures for Sign Language Recognition", *The IEEE International Conference on Computer Vision Workshops*, Seoul, South Korea, 2019.

157. Loshchilov, I. and F. Hutter, "Decoupled Weight Decay Regularization", *arXiv preprint arXiv:1711.05101*, 2017.

158. Gökçe, Ç., O. Özdemir, A. A. Kındıroğlu and L. Akarun, "Score-Level Multi Cue Fusion for Sign Language Recognition", *European Conference on Computer Vision*, pp. 294–309, Springer, Glassgow, UK, 2020.

159. Zhou, F. and F. De La Torre, "Generalized Canonical Time Warping", *Transactions on Pattern Analysis and Machine Intelligence*, Vol. 38, No. 2, pp. 279–294, 2015.

160. Okada, S., O. Aran and D. Gatica-Perez, "Personality Trait Classification via Co-Occurrent Multiparty Multimodal Event Discovery", *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction - ICMI '15*, pp. 15–22, ACM Press, New York, USA, 2015.

161. Kindiroglu, A. A., L. Akarun and O. Aran, "Vision Based Personality Analysis Using Transfer Learning Methods", *22nd Signal Processing and Communications Applications Conference (SIU)*, pp. 2050–2053, IEEE, Trabzon, Turkey, 2014.

162. Matthews, G., I. J. Deary and M. C. Whiteman, *Personality Traits*, Cambridge University Press, Cambridge, UK, 2003.

163. Vinciarelli, A. and G. Mohammadi, "A Survey of Personality Computing", *IEEE Transactions on Affective Computing*, Vol. 5, No. 3, pp. 273–291, 2014.

164. Sanchez-Cortes, D., O. Aran and D. Gatica-Perez, "An Audio Visual Corpus for Emergent Leader Analysis", *In Proceedings of Workshop on Multimodal corpora for Machine Learning: Taking Stock and Road Mapping the Future*, ICMI-MLMI, Alicante, Spain, 2011.

165. Biel, J. and D. Gatica-Perez, "The Youtube Lens: Crowdsourced Personality Impressions and Audiovisual Analysis of Vlogs", *Multimedia, IEEE Transactions on*, Vol. 1520-9210, pp. 41–55, 2013.

166. Alam, F., E. A. Stepanov and G. Riccardi, "Personality Traits Recognition on Social Network-Facebook", *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 7, pp. 6–9, Ann Arbor, USA, 2013.

167. Sanchez-Cortes, D., O. Aran, M. S. Mast and D. Gatica-Perez, "Identifying Emergent Leadership in Small Groups Using Nonverbal Communicative Cues", *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal interaction*, pp. 1–4, Beijing, China, 2010.

168. Gosling, S., P. Rentfrow and W. Swann, "A Very Brief Measure of the Big-Five Personality Domains", *Journal of Research in personality*, Vol. 37, No. 6, pp. 504–528, 2003.

169. Rammstedt, B. and O. P. John, "Measuring Personality in One Minute or Less: A 10-Item Short Version Of the Big Five Inventory in English and German", *Journal of Research in Personality*, Vol. 41, No. 1, pp. 203–212, 2007.

170. Mairesse, F., M. Walker, M. Mehl and R. Moore, "Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text", *Journal of artificial intelligence Research*, Vol. 30, No. 1, pp. 457–500, 2007.

171. Qiu, L., H. Lin, J. Ramsay and F. Yang, "You Are What You Tweet: Personality Expression and Perception on Twitter", *Journal of Research in Personality*, Vol. 46, No. 6, pp. 710–718, 2012.

172. Argamon, S., S. Dhawle, M. Koppel and J. W. Pennebaker, "Lexical Predictors Of Personality Type", *Proceedings of the Joint Annual Meeting of the Interface and Classification Society of North America*, St. Louis, USA, 2005.

173. Mohammadi, G. and A. Vinciarelli, "Automatic Personality Perception: Prediction of Trait Attribution Based on Prosodic Features", *IEEE Transactions on Affective Computing*, Vol. 3, No. 3, pp. 273–284, 2012.

174. Valente, F., S. Kim and P. Motlicek, "Annotation and Recognition of Personality Traits in Spoken Conversations from the AMI Meetings Corpus", *Proceedings of Interspeech 2012*, Portland, USA, 2012.

175. Fitzgerald Steele, David C. Evans, R. K. G., "Is Your Profile Picture Worth 1000 Words? Photo Characteristics Associated with Personality Impression Agreement.", *Proc. AAAI Int. Conf. Weblogs Social Media*, Minneapolis, USA, 2009.

176. Srivastava, R., J. Feng, S. Roy, S. Yan and T. Sim, "Don't Ask Me What I'm like, Just Watch and Listen", *Proceedings of the 20th ACM international Conference on Multimedia - MM '12*, p. 329, ACM Press, New York, USA, 2012.

177. Staiano, J., B. Lepri, R. Subramanian, N. Sebe and F. Pianesi, "Automatic Modeling of Personality States in Small Group Interactions", *Proceedings of the 19th ACM International Conference on Multimedia - MM '11*, p. 989, ACM Press, New York, USA, 2011.

178. Celli, F. and F. Pianesi, "Workshop on Computational Personality Recognition (shared Task)", *Proceedings of the Workshop on Computational Personality Recognition*, Orlando, USA, 2013.

179. Chittaranjan, G., J. Blom and D. Gatica-Perez, "Who's Who with Big-Five: Analyzing and Classifying Personality Traits with Smartphones", *2011 15th Annual International Symposium on Wearable Computers*, pp. 29–36, IEEE, San Francisco, USA, 2011.

180. Olguın, D., P. A. Gloor and A. S. Pentland, "Capturing Individual and Group Behavior with Wearable Sensors", *AAAI Spring Symposium on Human Behavior*

Modeling, Stanford, USA, 2009.

181. Yee, N., N. Ducheneaut, L. Nelson and P. Likarish, "Introverted Elves & Conscientious Gnomes", *Proceedings of the 2011 annual Conference on Human Factors in Computing Systems*, p. 753, ACM Press, New York, USA, 2011.

182. Ponce-López, V., B. Chen, M. Oliu, C. Corneanu, A. Clapés, I. Guyon, X. Baró, H. J. Escalante and S. Escalera, "ChaLearn LAP 2016: First Round Challenge on First Impressions-Dataset and Results", *European Conference on Computer Vision 2016 Workshops*, pp. 400–418, Springer, Amsterdam, Netherlands, 2016.

183. Escalante, H. J., V. Ponce-Lopez, J. Wan, M. A. Riegler, B. Chen, A. Clapes, S. Escalera, I. Guyon, X. Baro, P. Halvorsen, H. Müller and M. Larson, "ChaLearn Joint Contest on Multimedia Challenges Beyond Visual Analysis: An Overview", *International Conference on Pattern Recognition (ICPR 2016) Workshops*, Cancun, Mexico, 2016.

184. Gürpınar, F., H. Kaya and A. A. Salah, "Combining Deep Facial and Ambient Features for First Impression Estimation", *European Conference on Computer Vision 2016 Workshops*, pp. 372–385, Springer, Amsterdam, Netherlands, 2016.

185. Subramaniam, A., V. Patel, A. Mishra, P. Balasubramanian and A. Mittal, "Bi-Modal First Impressions Recognition Using Temporally Ordered Deep Audio and Stochastic Visual Features", *European Conference on Computer Vision 2016 Workshops*, pp. 337–348, Springer, Amsterdam, Netherlands, 2016.

186. Aydin, B., A. A. Kindiroglu, O. Aran and L. Akarun, "Automatic Personality Prediction from AudioVisual Data using Random Forest Regression", *International Conference on Pattern Recognition Workshops*, Cancun, Mexico, 2016.

187. Sanchez-Cortes, D., O. Aran, M. S. Mast and D. Gatica-Perez, "A Nonverbal Behavior Approach to Identify Emergent Leaders in Small Groups", *IEEE Trans-*

actions on Multimedia, Vol. 14, No. 3, pp. 816–832, 2012.

188. Aran, O. and D. Gatica-Perez, "One of A Kind: Inferring Personality Impressions in Meetings", *Proceedings of the 15th ACM on international Conference on Multimodal Interaction*, pp. 11–18, ACM Press, New York, USA, 2013.

189. Aran, O. and D. Gatica-Perez, "Cross-Domain Personality Prediction: From Video Blogs to Small Group Meetings", *Proceedings of the International Conference on Multimodal Interaction*, pp. 127–130, Sydney, Australia, 2013.

190. Guyon, I. and a. Elisseeff, "An Introduction to Variable and Feature Selection", *Journal of Machine Learning Research*, Vol. 3, pp. 1157–1182, 2003.

191. Chen, J., J. Liu and J. Ye, "Learning Incoherent Sparse and Low-Rank Patterns from Multiple Tasks", *ACM Transactions on Knowledge Discovery from Data (TKDD)*, Vol. 5, No. 4, pp. 1–31, 2012.

192. Zhang, Y. and D. Yeung, "Multi-Task Learning in Heterogeneous Feature Spaces", *Proceedings of the AAAI Conferrence on Artificial Intelligence*, San Francisco, USA, 2011.

193. Chang, C.-C. and C.-J. Lin, "LIBSVM: a Library for Support Vector Machines", *ACM Transactions on Intelligent Systems and Technology (TIST)*, Vol. 2, No. 3, pp. 1–27, 2011.

194. Teijeiro-Mosquera, L., J.-I. Biel, J. L. Alba-Castro and D. Gatica-Perez, "what Your Face Vlogs About: Expressions of Emotion and Big-Five Traits Impressions in You Tube", *IEEE Transactions on Affective Computing*, Vol. 6, No. 2, pp. 193–205, 2015.

195. Simonyan, K. and A. Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos", *arXiv preprint arXiv:1406.2199*, 2016.

196. Wang, H., X. Chai and X. Chen, "Sparse Observation (so) Alignment for Sign Language Recognition", *Neurocomputing*, Vol. 175, pp. 674–685, 2016.

197. Zhou, F. and F. de la Torre, "Canonical Time Warping for Alignment of Human Behavior", *Advances in Neural Information Processing Systems*, pp. 2286–2294, Vancouver, Canada, 2009.

198. Fu, A. W.-C., E. Keogh, L. Y. Lau, C. A. Ratanamahatana and R. C.-W. Wong, "Scaling and Time Warping in Time Series Querying", *The VLDB Journal—The International Journal on Very Large Data Bases*, Vol. 17, No. 4, pp. 899–921, 2008.

199. Tran, D., L. Bourdev, R. Fergus, L. Torresani and M. Paluri, "Learning Spatiotemporal Features with 3D Convolutional Networks", *Proceedings of the IEEE International Conference on Computer Vision*, Santiago, Chile, 2015.

200. Özdemir, O., N. C. Camgöz and L. Akarun, "Isolated Sign Language Recognition Using Improved Dense Trajectories", *2016 24th Signal Processing and Communication Application Conference (SIU)*, pp. 1961–1964, IEEE, Zonguldak, Turkey, 2016.

201. Camgoz, N. C., S. Hadfield, O. Koller and R. Bowden, "SubUNets: End-to-End Hand Shape and Continuous Sign Language Recognition", *International Conference on Computer Vision*, Vol. 2017-October, pp. 3075–3084, IEEE Computer Society, Venice, Italy, 2017.

202. Shi, L., Y. Zhang, J. Cheng and H. Lu, "Skeleton-Based Action Recognition with Multi-Stream Adaptive Graph Convolutional Networks", *IEEE Transactions on Image Processing*, Vol. 29, pp. 9532–9545, 2020.