Deep spatio-temporal structural model for free-form action recognition

Sephora Madjiheurem

Master Thesis
April 2016

Supervisors:
Jie Song
Dr. Mathieu Salzmann
Prof. Dr. Otmar Hilliges

ETH zürich

Advanced Interactive Technologies

Abstract

Human action is a high-level concept in computer vision research and understanding it may benefit from different semantics, such as human pose, interacting objects, and scene context.

In this thesis, we explicitly exploit semantic cues with aid of existing object detectors for action recognition in videos, and thoroughly study their effect on the recognition performance for different types of actions.

Specifically, we propose a new deep architecture by incorporating object/human detection results into the framework for action recognition.

Our proposed architecture not only shares great modeling capacity with two-stream input augmentation, but also exhibits the flexibility of leveraging semantic cues (e.g. scene, person, object) for action understanding.

We perform experiments on UCF101 dataset and demonstrate its superior performance to the original two-stream CNN. In addition, we systematically study the effect of incorporating semantic cues on the recognition performance for different types of action classes, and try to provide some insights for building more reasonable action benchmarks and robust recognition algorithms.

Ackowledgement

I would like to express my sincere gratitude to my supervisor Jie Song. His continuous guidance and boundless patience has given me immense inspiration and courage.

My sincere thanks also goes Dr.Limin Wang, whose insightful advice was crucial for the major breakthroughs in this thesis.

The completion of this thesis also owes to the tremendous love and care from my precious friends.

Especially, I would like to thank Srivathsan Murali, Seon-Wook Park for their endless encouragement and unconditioned help. Working beside you has been a great honour and your dedication to work has been a constant motivation for me.
To Pavol Vyhlidal, who taught me never cease to confront personal boundaries and take on new challenges in life.
To Federico Danieli, who has accompanied me through my ups and downs. Your valuable opinions gave rise to numerous progress; your loving support has been the greatest fuel that drove me forward.

Last but not least, I am thankful to my family. You enabled everything, EVERYTHING.

Contents

List of Figures

List of Tables

1 Introduction

2 Related Work

Human skeleton based action recognition

Spatio-temporal structures

3 Preliminary

3.1 Spatio-temporal structures

3.2 Deep Learning

Deep Learning is an area of Machine Learning that is characterized as a set of algorithms aiming at learning representations of data. These algorithms, known as Artificial Neural Networks (ANNs) are inspired by biological neural networks. An ANN is typically composed of multiple layers of artificial neurons, and each of these neural units are connected with many others. A deep learning model is able to extract from the data relevant features for solving a specific task. [REFERENCE TO MORE INFO ON DL]

Several deep learning architectures such as convolutional deep neural networks, deep belief networks and recurrent neural networks have led to state-of-the-art results on numerous applications in computer vision, natural language processing, audio recognition and bioinformatics [REFERENCES]

3.2.1 Recurrent Neural Networks

4 Methodology

4.1 Model Architecture

5 Implementation Details

5.1 Data

5.2 Training

5.3 Testing

6 Evaluation

In this chapter, we pursuit to answer the following three questions

1. Does semantic structure help action classification?
2. If so, how to bring out the best complementary effect?
3. How does semantic structure affect recognition performance for each action category?

In the following analysis, baseline refers to an in-house trained two-stream CNN network that does not integrate semantic cues. In other words, baseline model utilizes single scene cue. Without special declaration, evaluations in section 6.1~section 6.4 refers to UCF101 split1, while in section 6.5 we compare the classification performance of our proposed model with state-of-art methods across all UCF101 splits.

We choose train our own network instead of using a published model [7] because we have noticed that although the average classification accuracy stays consistent, the actual class accuracies have been shifted, meaning that the local minimum has slightly changed due to actual implementation differences. As we will cast a detailed investigation concerning category-wise performance in section 6.4, we use our own baseline in order to maintain a consistent local minimum for fair comparison.

The average accuracy of our baseline model and the published model is summarized in Table 6.1.

mAcc (%)	Split1		Split2		Split3		Avg	
Model	spatial	temporal	spatial	temporal	spatial	temporal	spatial	temporal
Theirs	79.8	85.7	77.3	88.2	77.8	87.4	78.4	87.0
Ours Baseline	79.42	85.27	77.14	88.13	77.25	86.96	77.93	86.79

Table 6.1: *Comparison of our in-house trained baseline with publicly available model [7]*

6.1 Contributions of Semantic Cues

In this section we verify whether incorporating explicit semantic structure defined by "person", "scene" and "object" cues enhances action recognition.

For this purpose we evaluate scene-only network and person-only network for spatial and temporal streams to study the relative importance of individual cues for different streams respectively. (Since object is regarded as an aiding cue, we omit evaluating its individual performance here.) To acquire an idea on how well scene and person cues complement each other, we apply a late-fusion by averaging the classification scores yielded from the two networks.

As we can see from Table 6.2, person and scene exhibit unequal relative relevance in spatial and temporal nets. While for spatial net, classification based on the whole scene is much more accurate than based on person (79.42% vs 73.82%), the relation is reversed in temporal net. This is not surprising since optical flow is inherently person-centric. On contrary, spatial net is fed on RGB information of single frames, where the appearance of action performer is usually less descriptive while the global contexts often provide essential hints for correct classification.

For both nets, combining person and scene in a late-fusion manner improves the classification accuracy, indicating the reciprocal property between the two cues.

From this point, we investigate the effect when incorporating both semantic cues into the same model as proposed in section 4.1 in favor of computational efficiency. As we can see from Table 6.2, the performance boost transfers to the proposed joint model and for spatial network the classification accuracy even exceeds that from late fusion (80.46% vs 80.10%).

Lastly, we incorporated object cue in spatial net (since object motion is too ambiguous in defining action, we omitted object cue for temporal net). Unfortunately, incooperating object channel does not generate significant improvement. While generating a marginal improvement compared to scene-only net (79.71% vs 79.42%), it falls behind "person+scene" model by 0.75%. We think the cause is manifold:

1. Object appearances have great intra-class variation as well as inter-class correlations. Particularly given that the object detections are still very noisy, the discriminant power in object channel is insufficient. In order to leverage object cue, a more expressive representation might be needed.

2. Since in the current implementation, we only utilize locational information of the object to extract sub-regions of the feature maps, object channel can be considered as an augmentation of scene channel. The summation of all three channels together implicitly puts more weight on scene channel, abating the strength of person channel.

3. The current implementation of MIL layer couldn't effectively update weights for relevant and irrelevant objects distinctively. As explained in section 4.1 object i's c-th entry (and the connected weights and lower layers) will be updated, as long as this object produces the strongest signal in class c among all objects in the input frame I. From our understanding, this algorithm can only work well under two assumptions: (1) there exist at least one relevant object in the frame (2) the detected objects are not relevant to other action classes. To elaborate this issue, consider we have detected two object regions containing "bow" and "green arena floor" in a frame from "Archery" action class. Assume

this frame is correctly classified, according to **??** a *positive* gradient will be passed to all other classes to *suppress* the corresponding classification signal. However, since "green arena floor" is the most representative region for classes such as "TennisSwing", it will be forced to decrease its response for "TennisSwing", although in reality it is positively contributive to this class.

mAcc(%)	S	P	S+P (two nets)	S+P (jointly)	S+P+O (jointly)
Spatial	79.42	73.82	80.16	**80.46**	79.71
Temporal	85.27	87.02	**87.84**	87.63	–

Table 6.2: *Integrating of explicit person cue (P) additionally to scene (S) is beneficial for both spatial and temporal streams. The contribution of object cue is marginal. Our proposed architecture (joint) is profitable in spatial net.*

6.2 Fusion Methods

From the previous section, we have learned that introducing person channel to action recognition effectively increases classification accuracy. In this section we investigate various fusion methods described in **??** so as to maximize the inter-channel complementary power. Considering training effort, all experiments are conducted using spatial net on split1 of UCF101 dataset.

The evaluation results are listed in Table 6.3.

First of all, opposed to our initial expectation, max fusion behaves considerably inadequately among all fusion variants. We believe this is due to the fact that since max operation only selectively updates the stronger channel (see **??**), it requires both channels to remain balanced in order to train all channels equally. During training if one channel appears to be stronger than the other, max would keep reinforcing the same channel, hence tilting the balance even further.

Secondly from Table 6.3 we also see that multiloss fusion variants yield inferior results than single-loss methods (except max). While multiloss performs well in jointly train a shared model for different tasks (e.g. regression and classification in Fast-RCNN [1]) and dataset augmentation as in [4], this scheme cannot sufficiently leverage the joint contribution from person and scene cues. According to our analysis, there are two possible reasons.

1. Unlike the aforementioned examples, where each sub-task is relatively independent, action recognition via different semantic cues are much stronger interplay, therefore sharing the same loss (thus a much similar parameter update in fc layers) could be beneficial for both channels;

2. Since scene region pooling always includes person region pooling and since person bounding boxes can contain noisy inputs, combining the classification scores helps recover individual input noise thus increase the robustness of the system.

Lastly, as is shown in Table 6.3 the two weighted fusion methods do not bring substantial improvement. We think this is because since during DNN training models always overfit to

training data, the inter-class confusion as well as relative cue importance cannot reflect the true distribution. Hence, weighting classification scores could not induce further performance boost. On contrary, increasing the total number of model parameters could lead to even severe overfitting.

Fusion	max	sum	weighted	cross weighted	multiloss (sum test)	multiloss (max test)	multiloss+
Avg Acc.(%)	78.95	**80.46**	79.77	80.20	79.15	79.03	78.91

Table 6.3: *Exploration of fusion methods using scene and person cues for spatial net on UCF101. Sum fusion outperforms other fusion methods.*

Based on previous empirical study, we are ready to build up our final action recognition architecture: sum-fused model with semantic channels "Scene" and "Person".

The final results on all 3 splits are summarized in Table 6.4.

mAcc (%)	Split1		Split2		Split3		Avg	
Stream	Baseline	Ours	Baseline	Ours	Baseline	Ours	Baseline	Ours
Spatial	79.42	**80.46**	**77.14**	76.53	77.25	**77.97**	77.93	**78.32**
Temporal (P)	85.27	**87.02**	88.13	**89.00**	86.96	**88.86**	86.79	**88.29**
Temporal (P+S)	85.27	**87.63**	88.13	**89.33**	86.96	**88.36**	86.79	**88.48**
Two Stream	90.98	**92.75**	91.45	**92.14**	91.05	**92.91**	91.15	**92.60**
Two Stream (Temp P+S)	90.98	**92.55**	91.45	**91.98**	91.05	**92.36**	91.15	**92.30**

Table 6.4: *Based on the previous empirical study, we propose using "scene" + "person" (P+S) model architecture for spatial and temporal network. We compare our proposed model with baseline over three splits on the UCF101 dataset, whereas baseline is the in-house implementation of two-stream CNNs. Two Stream results are yielded from summing spatial and temporal classification scores using weight* $1:3$.

6.3 ROI Quality

In this section we study the effect of the person detection quality on the model performance.

For this purpose, we focus on the annotated video dataset, JHMDB [2], and evaluate the spatial S+P model using (1) person ROIs from raw Faster-RCNN object detector, (2) filtered actor ROIs as described in **??** and (3) ground truth person ROIs.

	Baseline	S+P (raw)	S+P (filtered)	S+P (GT)
Accuracy (%)	51.16	52.01	53.77	54.25

Table 6.5: *Accuracy on JHMDB (split 1) using person ROIs of different quality. Similar as in UCF101, baseline refers to the inhouse trained model using solely "scene" channel.*

JHMDB is created from HMDB [3]. It is a fully annotated video dataset, which consists of 928 trimmed video clips from 21 action classes. These classes are mainly single-actor action classes,

As is shown in Table 6.5 the quality of extracted person ROIs directly affects accuracy for action recognition. For JHMDB dataset (mostly single person action), even raw detection results suffices to bring in an evident improvement; futhermore On the other hand, our filtered person ROIs are able to generate near groundtruth performance. This suggest that our method is robust against suboptimal ROI extraction.

6.4 Action Categories

In this section, we return to the action class categorization (**??**) proposed in chapter 1 that gave incentive to this thesis and evaluate the effectiveness of our model with respect to these categories.

Recall that we showed in **??** the baseline model performs significantly worse in action categories with weak dependence on scene. In our analysis earlier, this could be caused by the inefficiency of conventional method in autonomously learning to abstract the most discriminant semantic information from complex context. This becomes a much severe issue when the dataset inhabits small variance, as deep neural networks easily overfit to unrelated information.

Our proposed structure tackles this issue by providing explicit semantic structures. By design, it should be particularly useful for scene-independent classes. In this regard, we evaluate our model category-wise with respect to the baseline (scene only).

Figure 6.1 compares the per-category mean classification accuracy on split1 using different combinations of cues (we fix fusion method to be sum fusion, as it is the best performing one from the discussion in **??**). In Figure 6.2 we show the 10 most improved and most reduced action classes using our method compared to baseline.

To begin with, it should be noticed with increasing importance of scene, classification accuracy shows a clear rising trend both in spatial and temporal nets, which justifies our hypothesis that the conventional semantic-indifferent methods is sensitive to uninformative variability in scene channel.

Person Channel We first focus on the effect of person cue. As is shown in Figure 6.1, models with separate person channel introduce significant performance boost in *P* category for both spatial and flow nets. This proves that the integration of person cue is able to reduce the

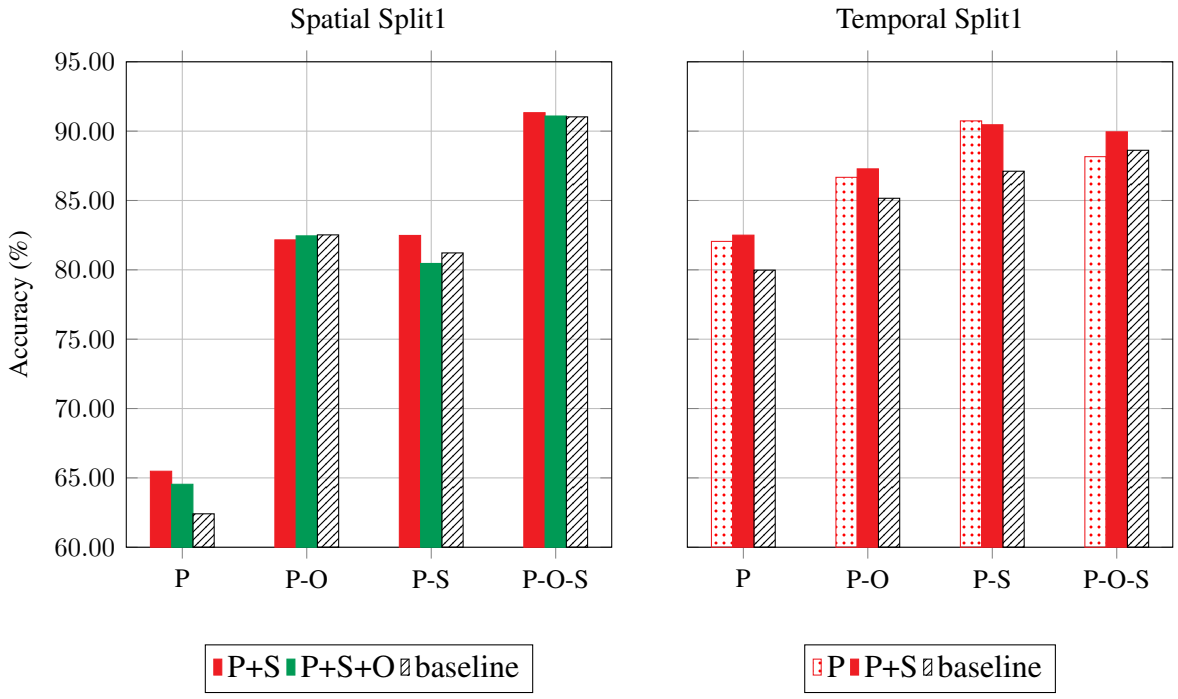


Figure 6.1: *Model performance evaluated according to different action categories on spatial stream (left) and flow stream (right). By integrating person cue, the performance of the originally weakest action category (motion only, P) is evidently enhanced.*

interference from scene channel.

Interestingly, an equally large improvement can be observed in *P-S* category. A careful examination over change of class-accuracies suggests that this improvement is mainly induced by action classes in similar scenes, for instance "FrontCrawl" vs "BreastStroke", "CricketBowling" vs "CricketShot" (ref Figure 6.2). This indicates, the utilization of person bounding boxes provides more refined pooling regions localizing the actual actions, which enables the identification of more sophisticated differences between similar actions.

While for spatial net the performance gain induced becomes less decisive in *P-O* and *P-O-S* categories, it remains striking for temporal net, suggesting that the person cue is especially beneficial for temporal net. This agrees with the conclusion we drew in section 6.1, namely for motion domain person is the dominating cue for successful action recognition, while for appearance domain the whole scene plays a more decisive role.

Object Channel On the other hand, the effect of object channel is not definite. As we can see from Figure 6.1, although the merged result improves the *P* category by a good margin (2.06% in Figure 6.1), this is in fact induced by person cue. In the remaining categories, object channel approximately aligns with scene channel, which confirms our assertion in section 6.1.

Complementarity Furthermore, Figure 6.3 depicts another noteworthy point. In this figure, we investigate the performance of individual channels *before* sum merging and their merged result in each action category. As we can see, person channel outperforms scene channel in

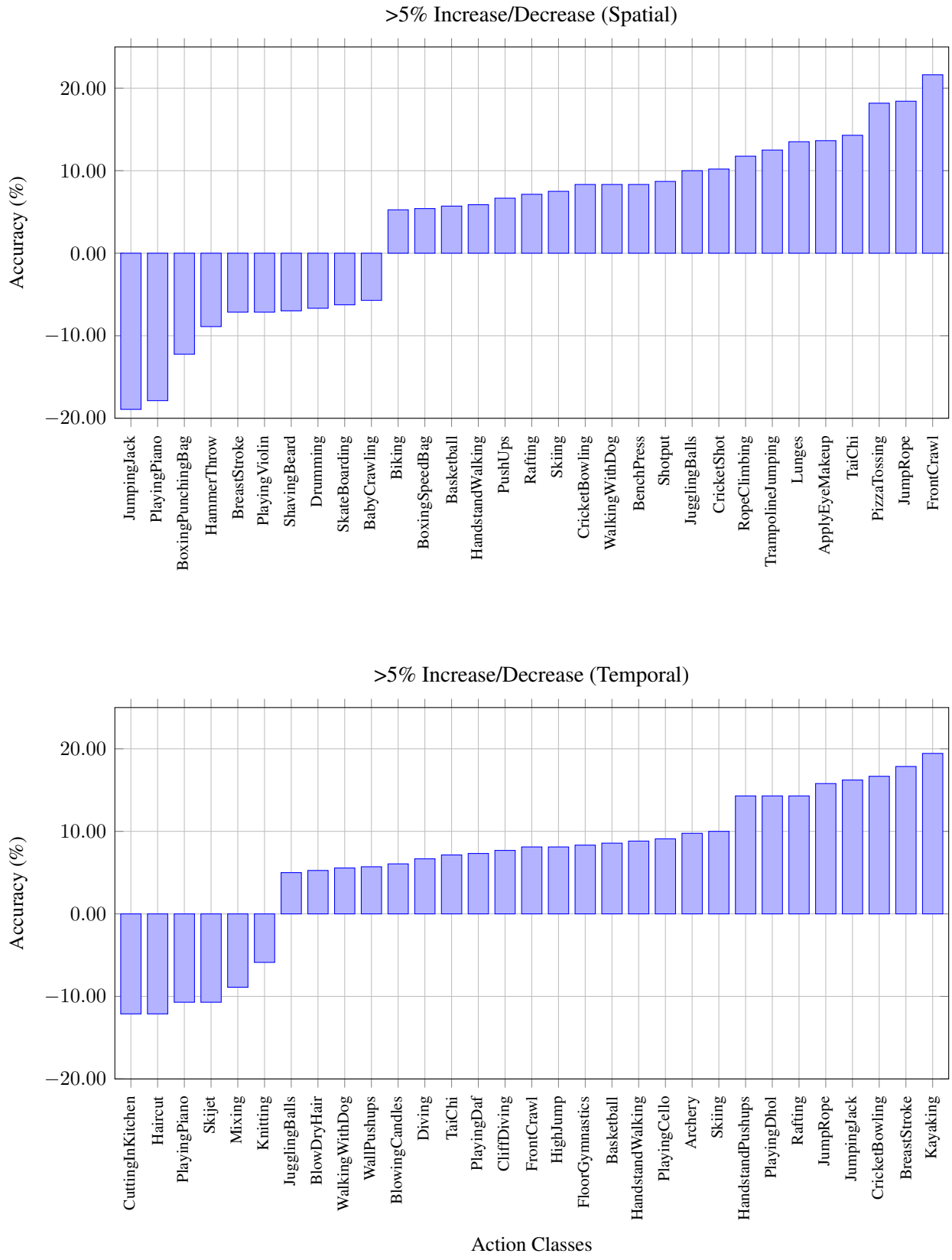


Figure 6.2: *Classes with $> 5\%$ performance gain and decline in spatial and temporal using our proposed network (P+S).*

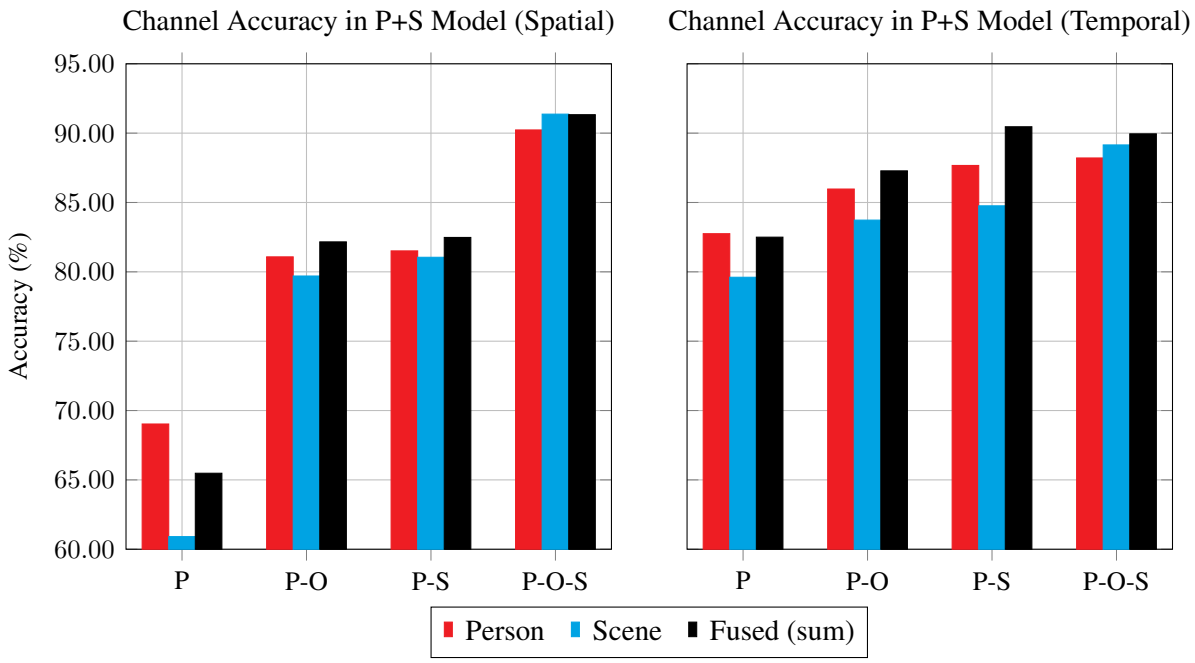


Figure 6.3: *Performance from individual channels before merging unit in spatial network with sum-fused person and scene channels. The relative performance discrepancy echoes our action class categorization. The merged aligns with the stronger channel, showing the complementary effects of our model.*

most action categories. As the scene becomes more indicative, the advantage of person channel gradually resides until it is overtaken in *P-O-S* category, where the scene is overwhelmingly indicative. This trends confirms our action categorization. At the same time, it implies that when explicitly decomposed either channel is able to unfold its classification strength in its own "specialization".

Moreover, as is black plot in Figure 6.3 shows, the merged results either align with (in *P-O*, *P-S* and *P-O-S* categories) or evidently is boosted (in *P* category) by the stronger channel. This implies that our proposed model is able to exploit the strength from both semantic channels in an effective and complementary way.

6.5 Comparison with State-of-the-art

Finally we compare our proposed approach to other state-of-the-art methods on UCF101 dataset.

The results are summarized in Table 6.6, where we compare our result with both traditional approaches such as improved trajectories (iDTs), and deep learning representations, such as Two Stream and Deep Two Stream.

Method	iDT + FV [6]	C3D [5]	TwoStream [4]	TwoStream [4] + LSTM [8]	Deep TwoStream [7]	Ours
Avg.Acc (%)	85.9	85.2	88.0	88.6	91.4	**92.6**

Table 6.6: *Comparison with the state-of-the-art methods on the UCF101 dataset.*

Conclusion and Future Work

7.1 Conclusion

7.2 Future Work

Appendix

A

A.1 Categorization of UCF101

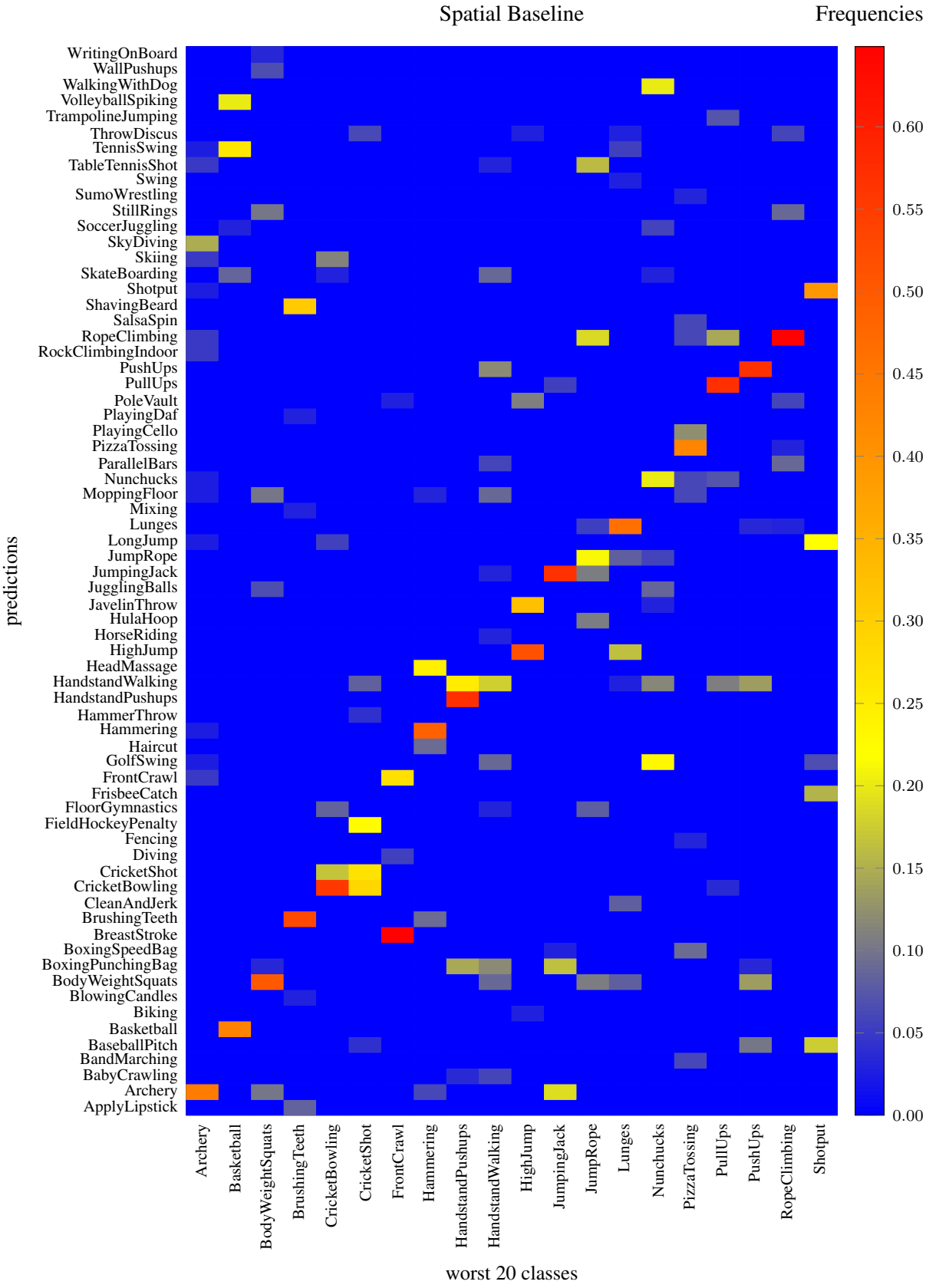
Body Motion (*P*)	ApplyEyeMakeup, ApplyLipstick, BabyCrawling, Basketball, BodyWeightSquats, BoxingSpeedBag, BrushingTeeth, Haircut, HandstandPushups, HandstandWalking, HeadMassage, HulaHoop, JugglingBalls, JumpingJack, JumpRope, Lunges, PullUps, PushUps, RopeClimbing, SalsaSpin, ShavingBeard, TaiChi, WallPushups, YoYo
Human-Object Interaction (*P-O*)	Archery, BenchPress, Biking, BlowDryHair, BlowingCandles, BoxingPunchingBag, CleanAndJerk, Drumming, Hammering, HorseRiding, Knitting, Mixing, Nunchucks, PizzaTossing, PlayingCello, PlayingDaf, PlayingDhol, PlayingFlute, PlayingGuitar, PlayingPiano, PlayingSitar, PlayingTabla, PlayingViolin, SoccerJuggling, Typing, WalkingWithDog
Body Motion in specific Scene (*P-S*)	BandMarching, BasketballDunk, BreastStroke, CliffDiving, CricketBowling, CricketShot, Diving, Fencing, FieldHockeyPenalty, FloorGymnastics, FrisbeeCatch, FrontCrawl, GolfSwing, HammerThrow, HighJump, IceDancing, JavelinThrow, LongJump, MilitaryParade, Punch, RockClimbingIndoor, Shotput, SkyDiving, SumoWrestling, Surfing, ThrowDiscus, VolleyballSpiking
Human-Object Interaction in specific Scene (*P-O-S*)	BalanceBeam, BaseballPitch, Billiards, Bowling, CuttingInKitchen, HorseRace, Kayaking, MoppingFloor, ParallelBars, PoleVault, PommelHorse, Rafting, Rowing, SkateBoarding, Skiing, Skijet, SoccerPenalty, StillRings, Swing, TableTennisShot, TennisSwing, TrampolineJumping, UnevenBars, WritingOnBoard

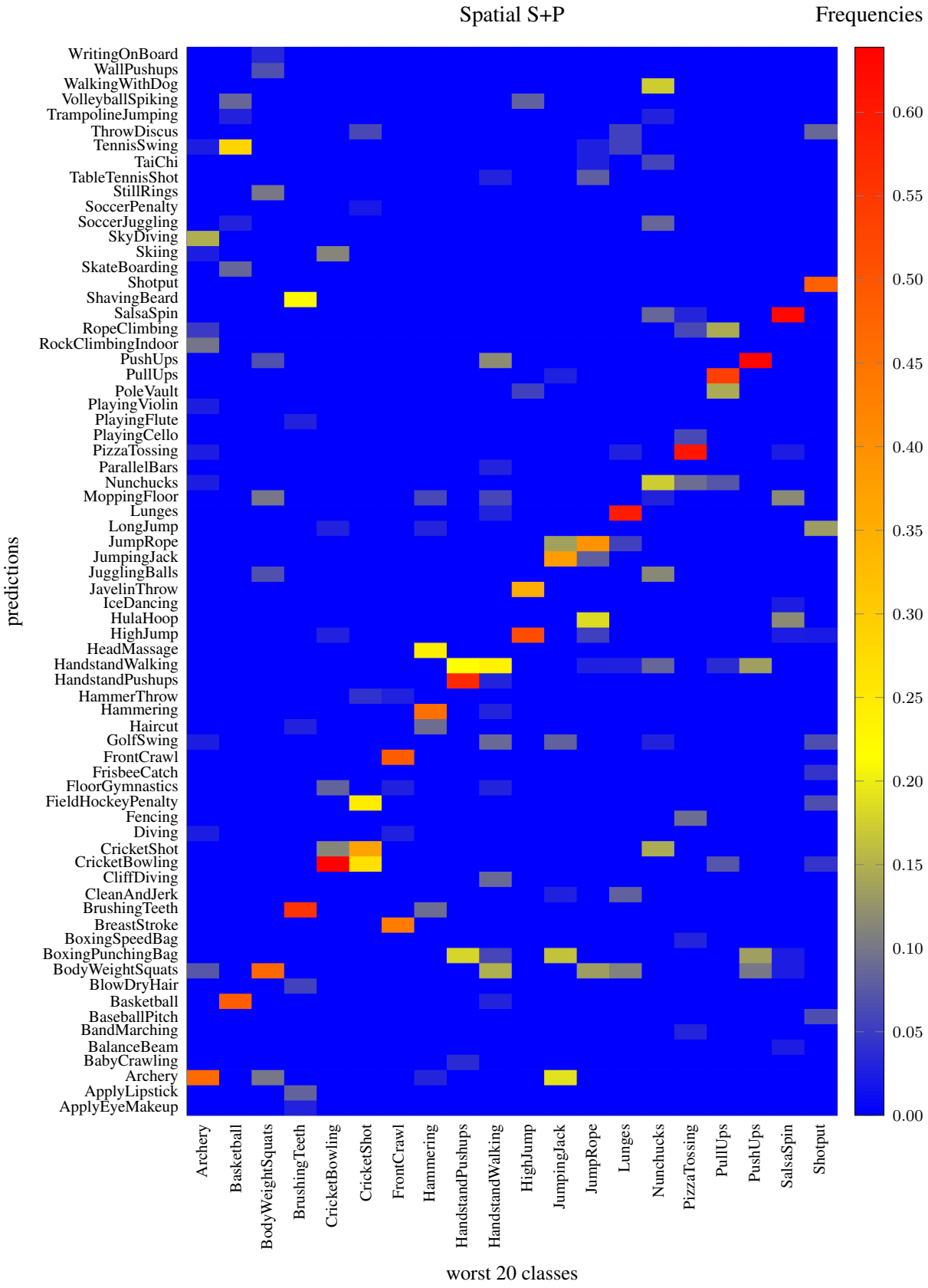
Table A.1: *Categorization of action classes in UCF101 dataset according to their semantic composition.*

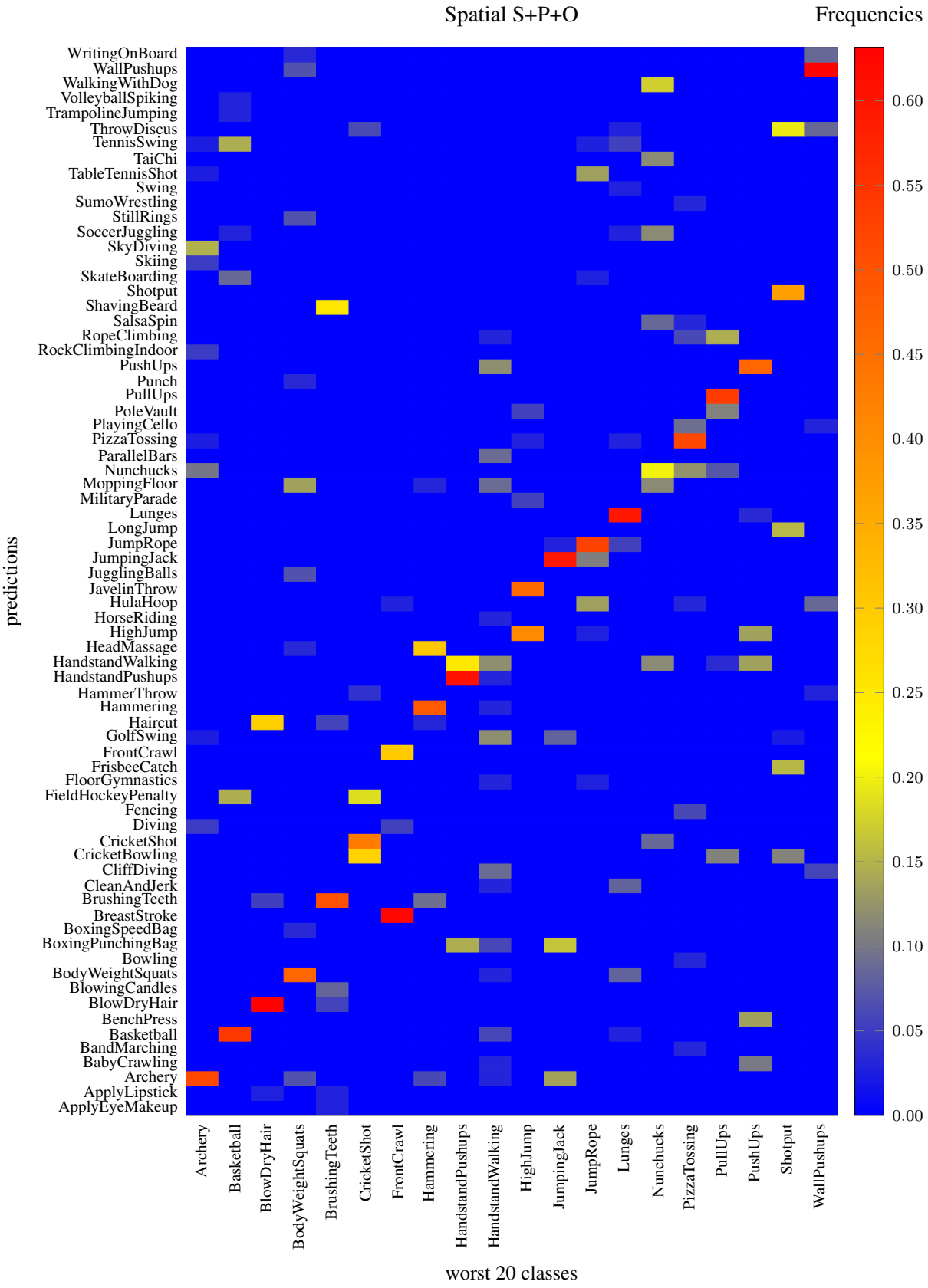
A.2 Object Categories in Faster-RCNN

id	Name	id	Name	id	Name	id	Name
1	accordion	34	dishwasher	66	nail	98	stove
2	airplane	35	drum	67	neck brace	99	sunglasses
3	axe	36	dumbbell	68	person	100	swimming trunks
4	baby bed	37	electric fan	69	piano	101	table
5	backpack	38	face powder	70	pineapple	102	tennis ball
6	balance beam	39	flower pot	71	ping-pong ball	103	tie
7	band aid	40	frying pan	72	pizza	104	toaster
8	baseball	41	golf ball	73	plastic bag	105	traffic light
9	basketball	42	golfcart	74	pomegranate	106	train
10	bathing cap	43	guitar	75	popsicle	107	trombone
11	beaker	44	hair dryer	76	power drill	108	trumpet
12	bench	45	hair spray	77	pretzel	109	tv or monitor
13	bicycle	46	hamburger	78	printer	110	unicycle
14	bookshelf	47	hammer	79	puck	111	vacuum
15	bow	48	harmonica	80	punching bag	112	violin
16	bowl	49	harp	81	purse	113	volleyball
17	brassiere	50	hat with a wide brim	82	racket	114	washer
18	bus	51	helmet	83	refrigerator	115	water bottle
19	can opener	52	horizontal bar	84	remote control	116	watercraft
20	car	53	horse	85	rubber eraser	117	wine bottle
21	cart	54	iPod	86	rugby ball	118	bottle
22	cello	55	ladle	87	ruler		
23	chain saw	56	lamp	88	salt or pepper shaker		
24	chair	57	laptop	89	saxophone		
25	cocktail shaker	58	lipstick	90	screwdriver		
26	coffee maker	59	maillot	91	ski		
27	computer keyboard	60	microphone	92	snowmobile		
28	computer mouse	61	microwave	93	snowplow		
29	corkscrew	62	milk can	94	soap dispenser		
30	croquet ball	63	miniskirt	95	soccer ball		
31	cup or mug	64	motorcycle	96	sofa		
32	diaper	65	mushroom	97	stethoscope		
33	digital clock						

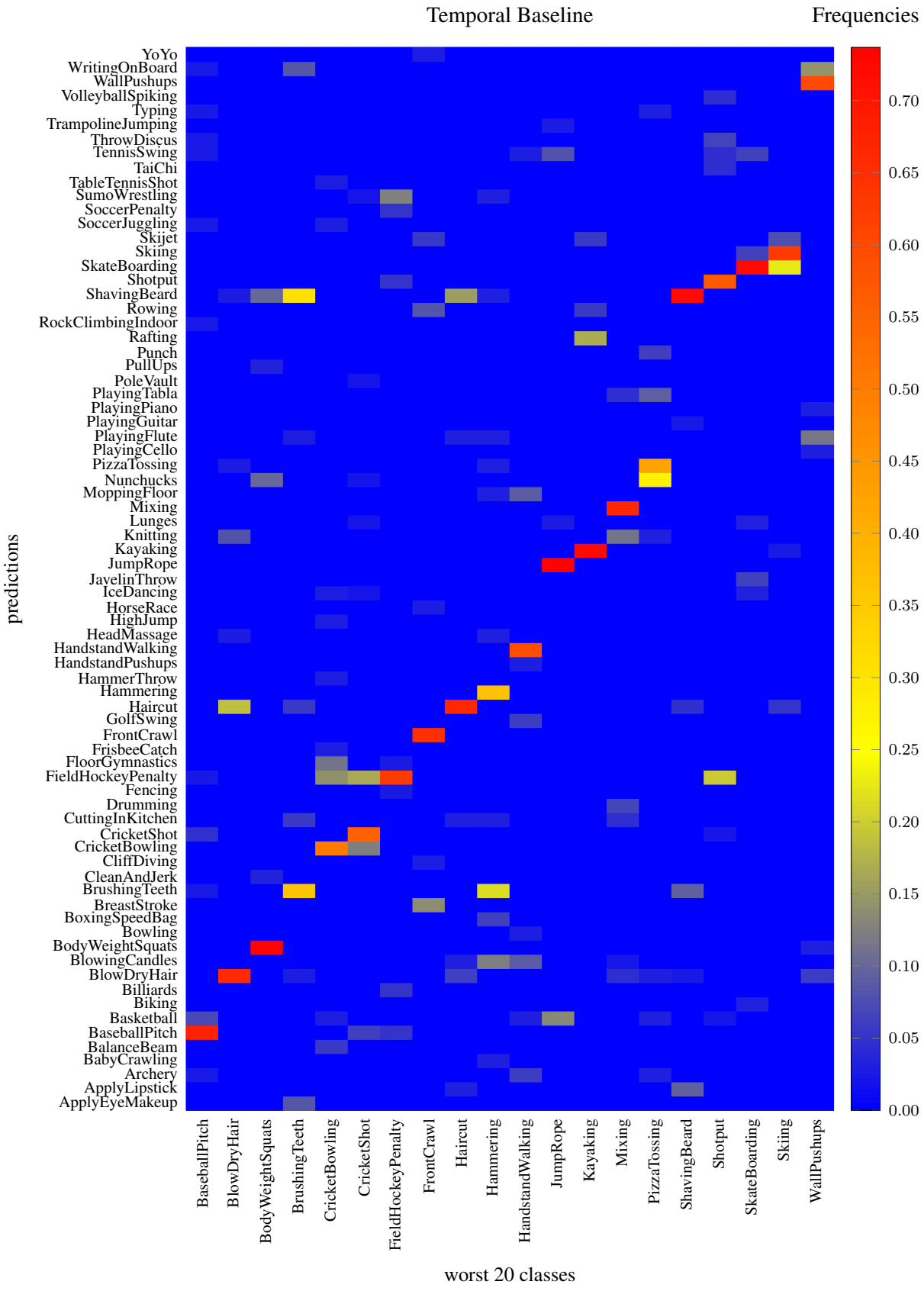
A.3 Confusion Maps

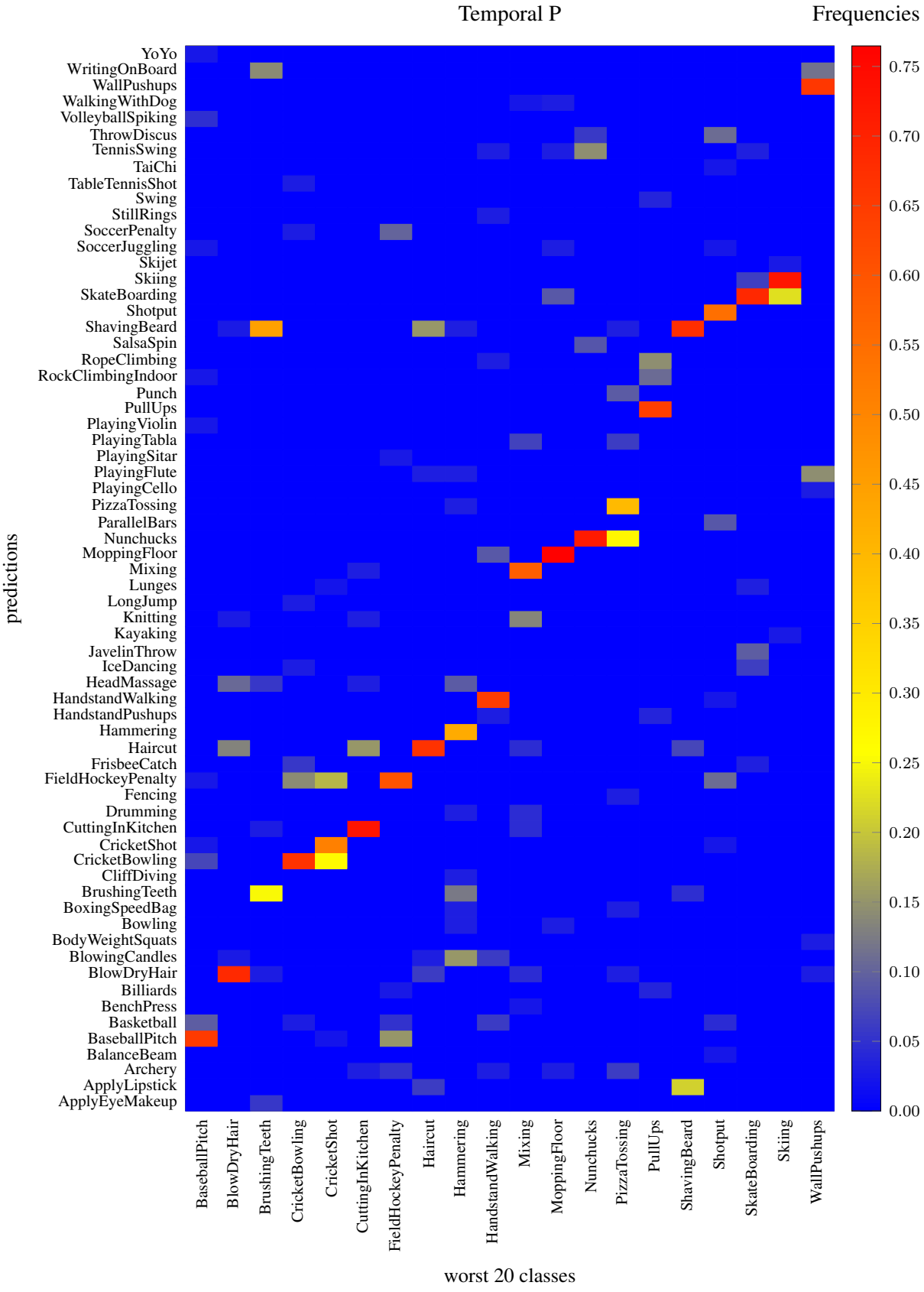



Spatial S+P
Frequencies
predictions
WritingOnBoard
WallPushups
WalkingWithDog
VolleyballSpiking
TrampolineJumping
ThrowDiscus
TennisSwing
TaiChi
TableTennisShot
StillRings
SoccerPenalty
SoccerJuggling
SkyDiving
Skiing
SkateBoarding
Shotput
ShavingBeard
SalsaSpin
RopeClimbing
RockClimbingIndoor
PushUps
PullUps
PoleVault
PlayingViolin
PlayingFlute
PlayingCello
PizzaTossing
ParallelBars
Nunchucks
MoppingFloor
Lunges
LongJump
JumpRope
JumpingJack
JugglingBalls
JavelinThrow
IceDancing
HulaHoop
HighJump
HeadMassage
HandstandWalking
HandstandPushups
HammerThrow
Hammering
Haircut
GolfSwing
FrontCrawl
FrisbeeCatch
FloorGymnastics
FieldHockeyPenalty
Fencing
Diving
CricketShot
CricketBowling
CliffDiving
CleanAndJerk
BrushingTeeth
BreastStroke
BoxingSpeedBag
BoxingPunchingBag
BodyWeightSquats
BlowDryHair
Basketball
BaseballPitch
BandMarching
BalanceBeam
BabyCrawling
Archery
ApplyLipstick
ApplyEyeMakeup
Archery
Basketball
BodyWeightSquats
BrushingTeeth
CricketBowling
CricketShot
FrontCrawl
Hammering
HandstandPushups
HandstandWalking
HighJump
JumpingJack
JumpRope
Lunges
Nunchucks
PizzaTossing
PullUps
PushUps
SalsaSpin
Shotput
worst 20 classes
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00

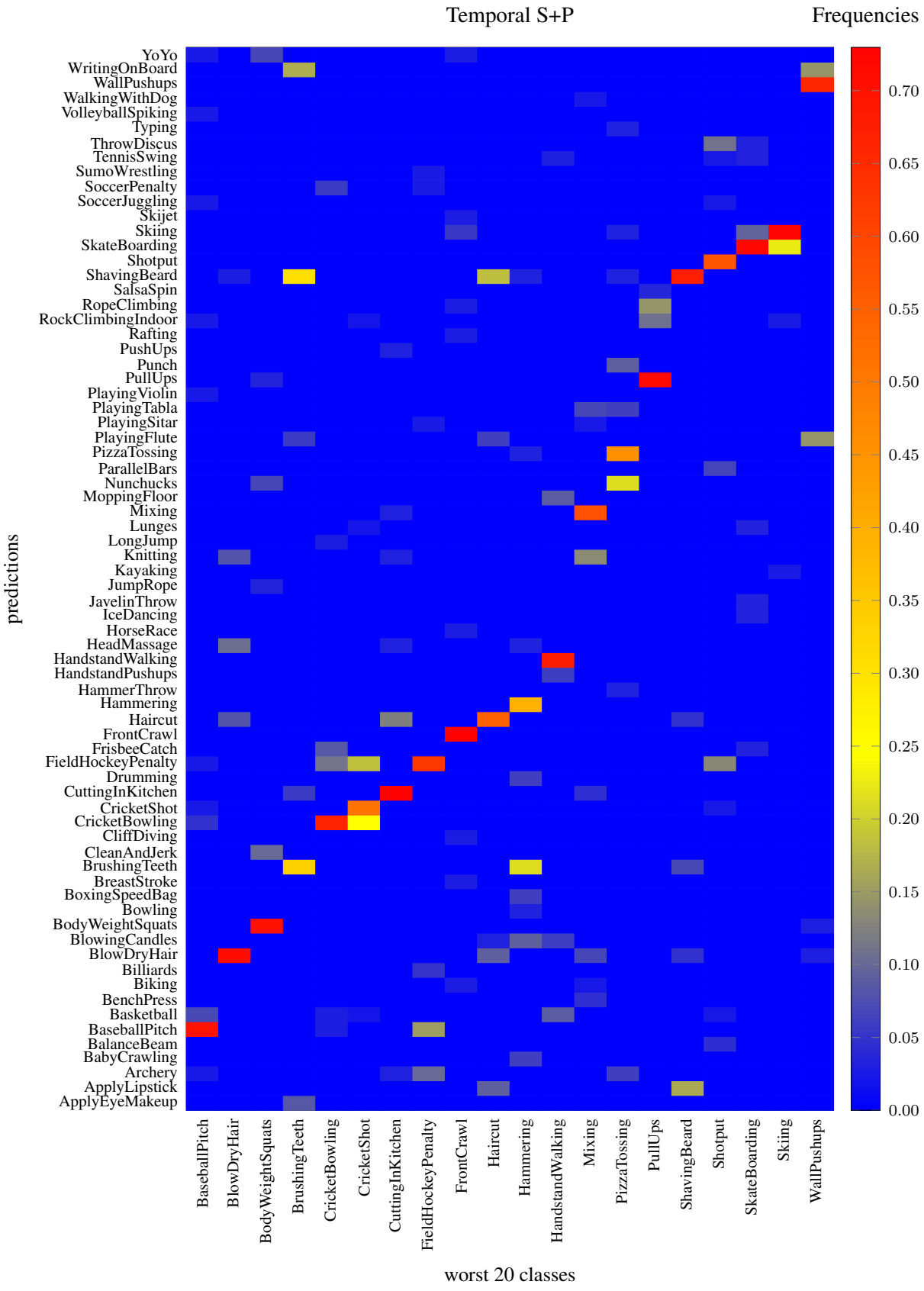

Spatial S+P+O
Frequencies
predictions
WritingOnBoard
WallPushups
WalkingWithDog
VolleyballSpiking
TrampolineJumping
ThrowDiscus
TennisSwing
TaiChi
TableTennisShot
Swing
SumoWrestling
StillRings
SoccerJuggling
SkyDiving
Skiing
SkateBoarding
Shotput
ShavingBeard
SalsaSpin
RopeClimbing
RockClimbingIndoor
PushUps
Punch
PullUps
PoleVault
PlayingCello
PizzaTossing
ParallelBars
Nunchucks
MoppingFloor
MilitaryParade
Lunges
LongJump
JumpRope
JumpingJack
JugglingBalls
JavelinThrow
HulaHoop
HorseRiding
HighJump
HeadMassage
HandstandWalking
HandstandPushups
HammerThrow
Hammering
Haircut
GolfSwing
FrontCrawl
FrisbeeCatch
FloorGymnastics
FieldHockeyPenalty
Fencing
Diving
CricketShot
CricketBowling
CliffDiving
CleanAndJerk
BrushingTeeth
BreastStroke
BoxingSpeedBag
BoxingPunchingBag
Bowling
BodyWeightSquats
BlowingCandles
BlowDryHair
BenchPress
Basketball
BandMarching
BabyCrawling
Archery
ApplyLipstick
ApplyEyeMakeup
Archery
Basketball
BlowDryHair
BodyWeightSquats
BrushingTeeth
CricketShot
FrontCrawl
Hammering
HandstandPushups
HandstandWalking
HighJump
JumpingJack
JumpRope
Lunges
Nunchucks
PizzaTossing
PullUps
PushUps
Shotput
WallPushups
worst 20 classes
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00

A. Appendix


Temporal Baseline
Frequencies
predictions
YoYo
WritingOnBoard
WallPushups
VolleyballSpiking
Typing
TrampolineJumping
ThrowDiscus
TennisSwing
TaiChi
TableTennisShot
SumoWrestling
SoccerPenalty
SoccerJuggling
Skijet
Skiing
SkateBoarding
Shotput
ShavingBeard
Rowing
RockClimbingIndoor
Rafting
Punch
PullUps
PoleVault
PlayingTabla
PlayingPiano
PlayingGuitar
PlayingFlute
PlayingCello
PizzaTossing
Nunchucks
MoppingFloor
Mixing
Lunges
Knitting
Kayaking
JumpRope
JavelinThrow
IceDancing
HorseRace
HighJump
HeadMassage
HandstandWalking
HandstandPushups
HammerThrow
Hammering
Haircut
GolfSwing
FrontCrawl
FrisbeeCatch
FloorGymnastics
FieldHockeyPenalty
Fencing
Drumming
CuttingInKitchen
CricketShot
CricketBowling
CliffDiving
CleanAndJerk
BrushingTeeth
BreastStroke
BoxingSpeedBag
Bowling
BodyWeightSquats
BlowingCandles
BlowDryHair
Billiards
Biking
Basketball
BaseballPitch
BalanceBeam
BabyCrawling
Archery
ApplyLipstick
ApplyEyeMakeup
BaseballPitch
BlowDryHair
BodyWeightSquats
BrushingTeeth
CricketBowling
CricketShot
FieldHockeyPenalty
FrontCrawl
Haircut
Hammering
HandstandWalking
JumpRope
Kayaking
Mixing
PizzaTossing
ShavingBeard
Shotput
SkateBoarding
Skiing
WallPushups
worst 20 classes
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00


Temporal P
Frequencies
predictions
YoYo
WritingOnBoard
WallPushups
WalkingWithDog
VolleyballSpiking
ThrowDiscus
TennisSwing
TaiChi
TableTennisShot
Swing
StillRings
SoccerPenalty
SoccerJuggling
Skijet
Skiing
SkateBoarding
Shotput
ShavingBeard
SalsaSpin
RopeClimbing
RockClimbingIndoor
Punch
PullUps
PlayingViolin
PlayingTabla
PlayingSitar
PlayingFlute
PlayingCello
PizzaTossing
ParallelBars
Nunchucks
MoppingFloor
Mixing
Lunges
LongJump
Knitting
Kayaking
JavelinThrow
IceDancing
HeadMassage
HandstandWalking
HandstandPushups
Hammering
Haircut
FrisbeeCatch
FieldHockeyPenalty
Fencing
Drumming
CuttingInKitchen
CricketShot
CricketBowling
CliffDiving
BrushingTeeth
BoxingSpeedBag
Bowling
BodyWeightSquats
BlowingCandles
BlowDryHair
Billiards
BenchPress
Basketball
BaseballPitch
BalanceBeam
Archery
ApplyLipstick
ApplyEyeMakeup
BaseballPitch
BlowDryHair
BrushingTeeth
CricketBowling
CricketShot
CuttingInKitchen
FieldHockeyPenalty
Haircut
Hammering
HandstandWalking
Mixing
MoppingFloor
Nunchucks
PizzaTossing
PullUps
ShavingBeard
Shotput
SkateBoarding
Skiing
WallPushups
worst 20 classes
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00

A. Appendix


Temporal S+P
Frequencies
predictions
YoYo
WritingOnBoard
WallPushups
WalkingWithDog
VolleyballSpiking
Typing
ThrowDiscus
TennisSwing
SumoWrestling
SoccerPenalty
SoccerJuggling
Skijet
Skiing
SkateBoarding
Shotput
ShavingBeard
SalsaSpin
RopeClimbing
RockClimbingIndoor
Rafting
PushUps
Punch
PullUps
PlayingViolin
PlayingTabla
PlayingSitar
PlayingFlute
PizzaTossing
ParallelBars
Nunchucks
MoppingFloor
Mixing
Lunges
LongJump
Knitting
Kayaking
JumpRope
JavelinThrow
IceDancing
HorseRace
HeadMassage
HandstandWalking
HandstandPushups
HammerThrow
Hammering
Haircut
FrontCrawl
FrisbeeCatch
FieldHockeyPenalty
Drumming
CuttingInKitchen
CricketShot
CricketBowling
CliffDiving
CleanAndJerk
BrushingTeeth
BreastStroke
BoxingSpeedBag
Bowling
BodyWeightSquats
BlowingCandles
BlowDryHair
Billiards
Biking
BenchPress
Basketball
BaseballPitch
BalanceBeam
BabyCrawling
Archery
ApplyLipstick
ApplyEyeMakeup
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
BaseballPitch
BlowDryHair
BodyWeightSquats
BrushingTeeth
CricketBowling
CricketShot
CuttingInKitchen
FieldHockeyPenalty
FrontCrawl
Haircut
Hammering
HandstandWalking
Mixing
PizzaTossing
PullUps
ShavingBeard
Shotput
SkateBoarding
Skiing
WallPushups
worst 20 classes

Bibliography

[1] Ross Girshick, *Fast r-cnn*, Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1440–1448.

[2] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael Black, *Towards understanding action recognition*, Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 3192–3199.

[3] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, *HMDB: a large video database for human motion recognition*, Proceedings of the International Conference on Computer Vision (ICCV), 2011.

[4] Karen Simonyan and Andrew Zisserman, *Two-stream convolutional networks for action recognition in videos*, Advances in Neural Information Processing Systems, 2014, pp. 568–576.

[5] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, *Learning spatiotemporal features with 3d convolutional networks*, ICCV, 2015, pp. 4489–4497.

[6] Heng Wang and Cordelia Schmid, *Action recognition with improved trajectories*, Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 3551–3558.

[7] Limin Wang, Yuanjun Xiong, Zhe Wang, and Yu Qiao, *Towards good practices for very deep two-stream convnets*, arXiv preprint arXiv:1507.02159 (2015).

[8] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici, *Beyond short snippets: Deep networks for video classification*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4694–4702.