

# 1 Categorization of UCF101

Body Motion ( <i>P</i> )	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 ( <i>P-O</i> )	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 ( <i>P-S</i> )	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 ( <i>P-O-S</i> )	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 1: Categorization of action classes in UCF101 dataset according to their semantic composition.

## 2 Object Categories in Faster-RCNN

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

### 3 Most Improved and Declined Action Classes

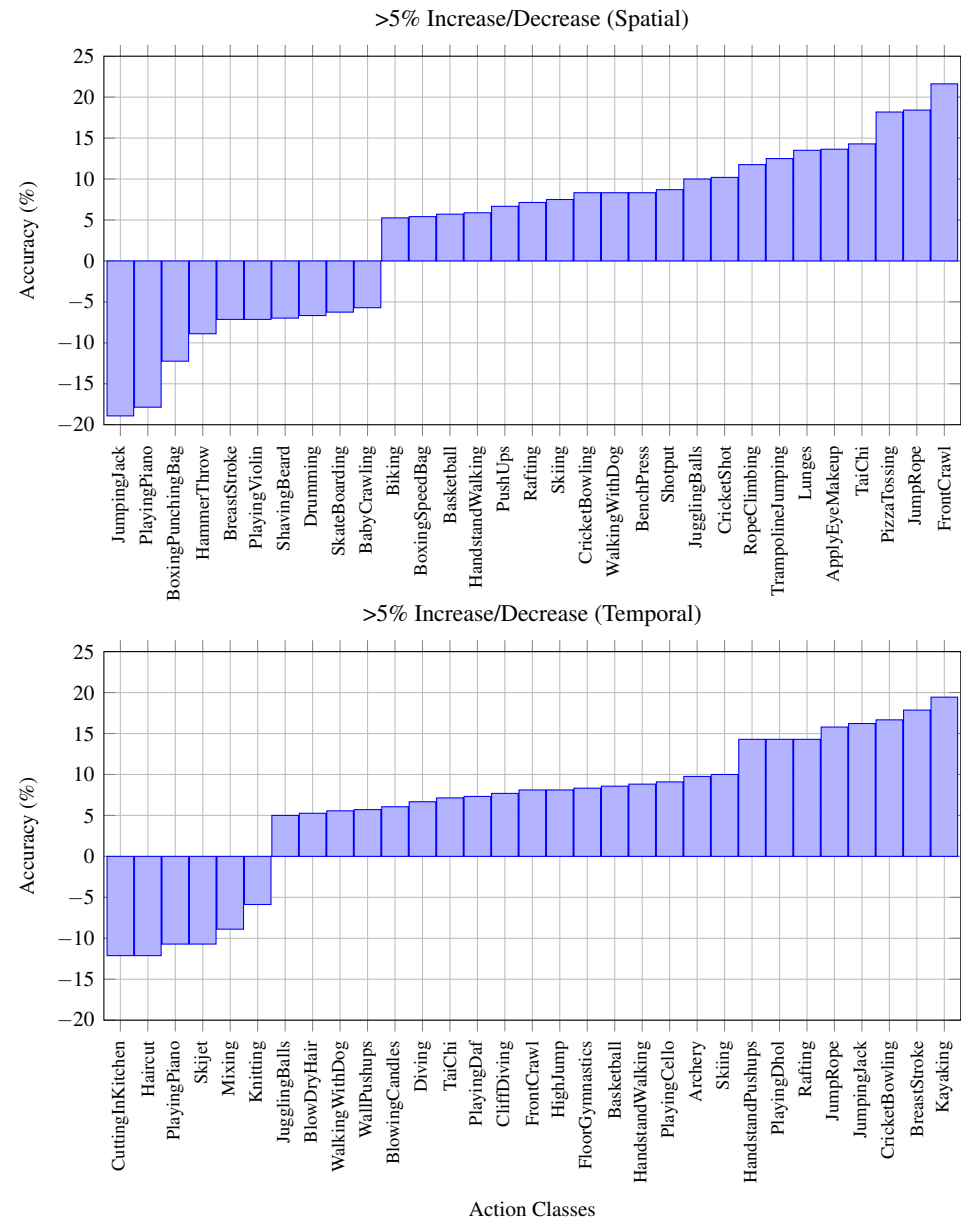


Figure 1: UCF101 action classes with > 5% performance gain or decline in spatial and temporal streams using our proposed network (P+S).

## 4 Confusion Maps

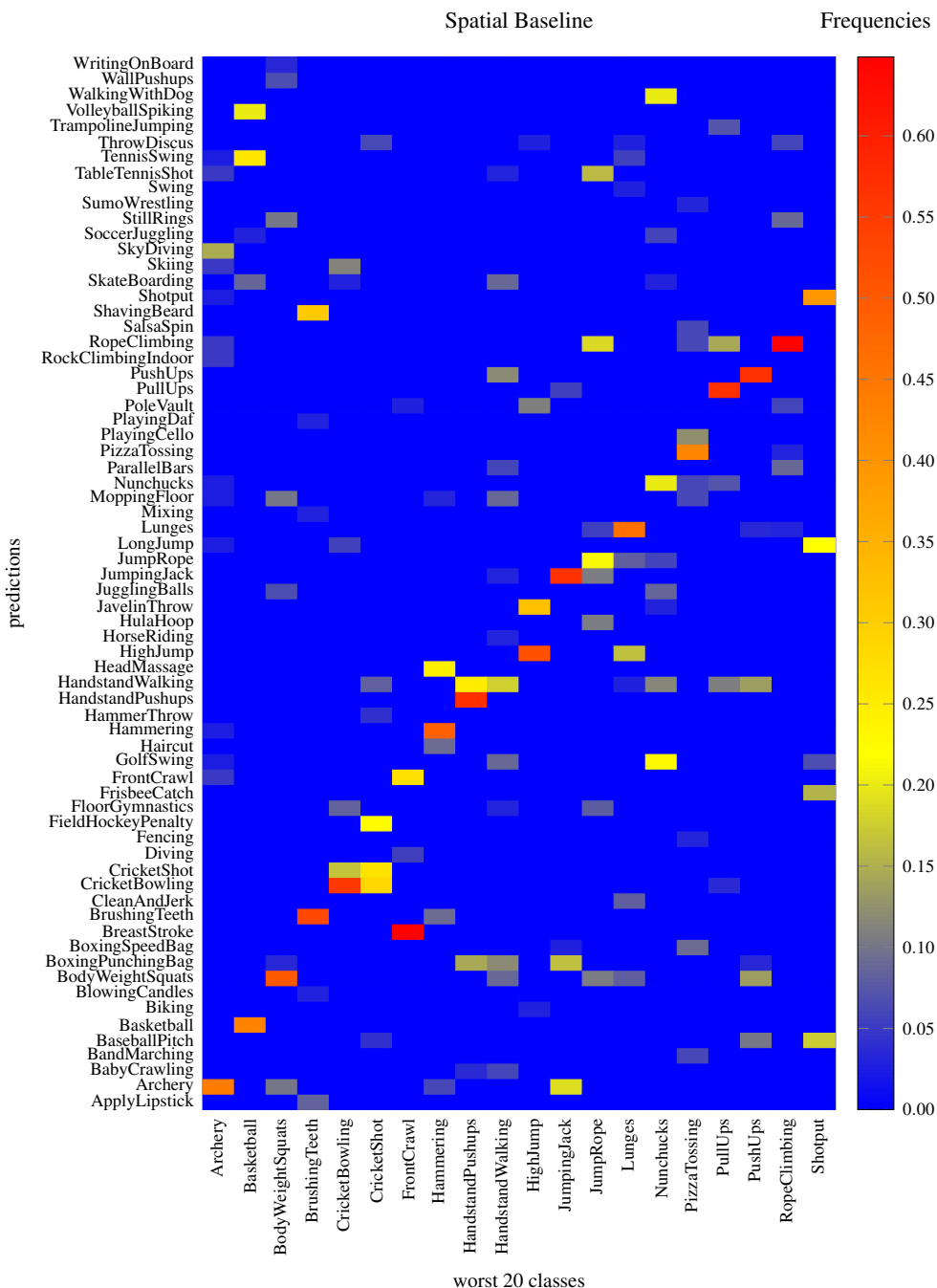


Figure 2: Confusions of the 20 worst performing action classes in spatial stream using base-line ("scene" only) model.

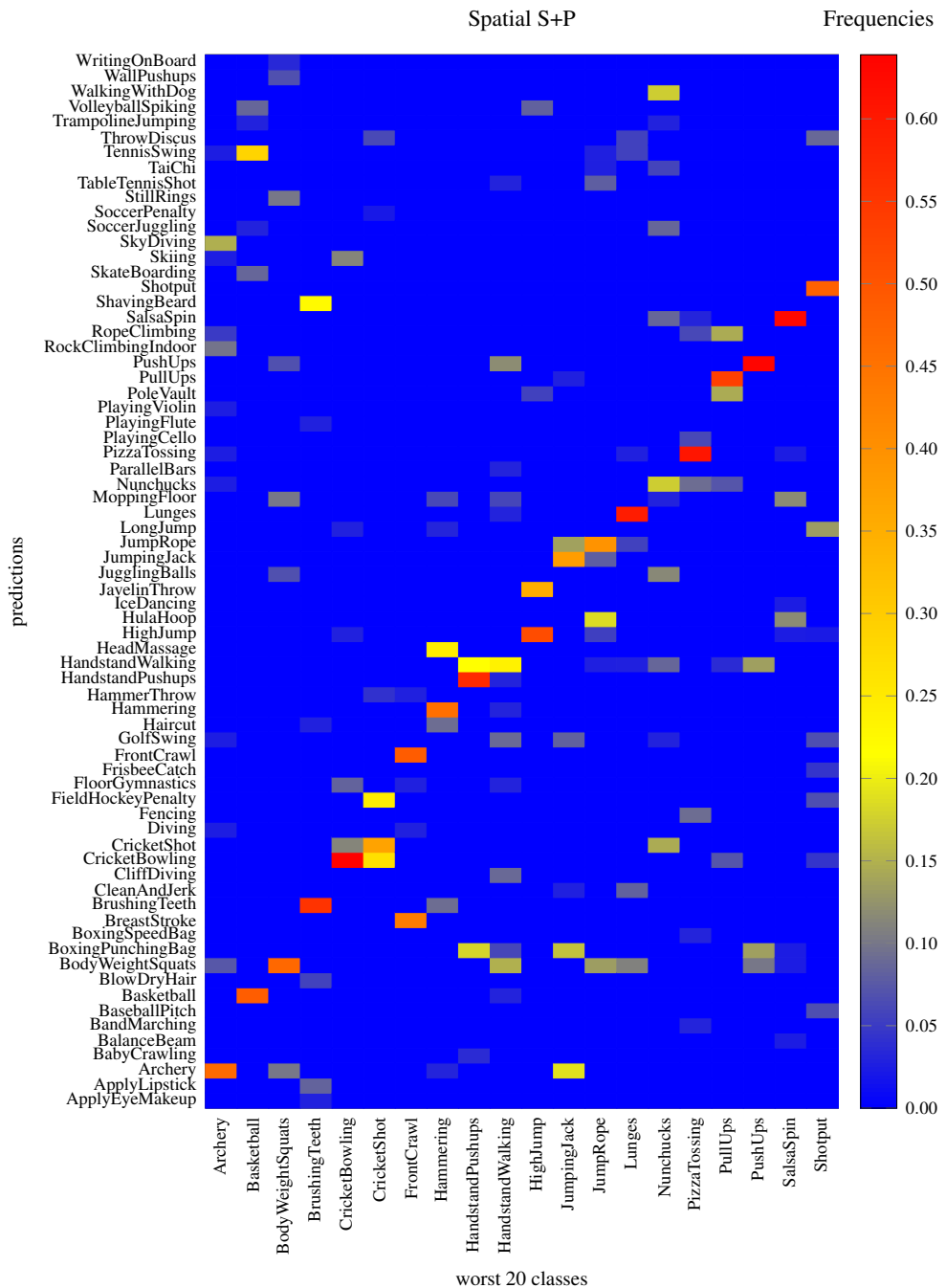


Figure 3: Confusions of the 20 worst performing action classes in spatial stream using "scene" and "person" cues (S+P model).

Spatial S+P+O

Frequencies

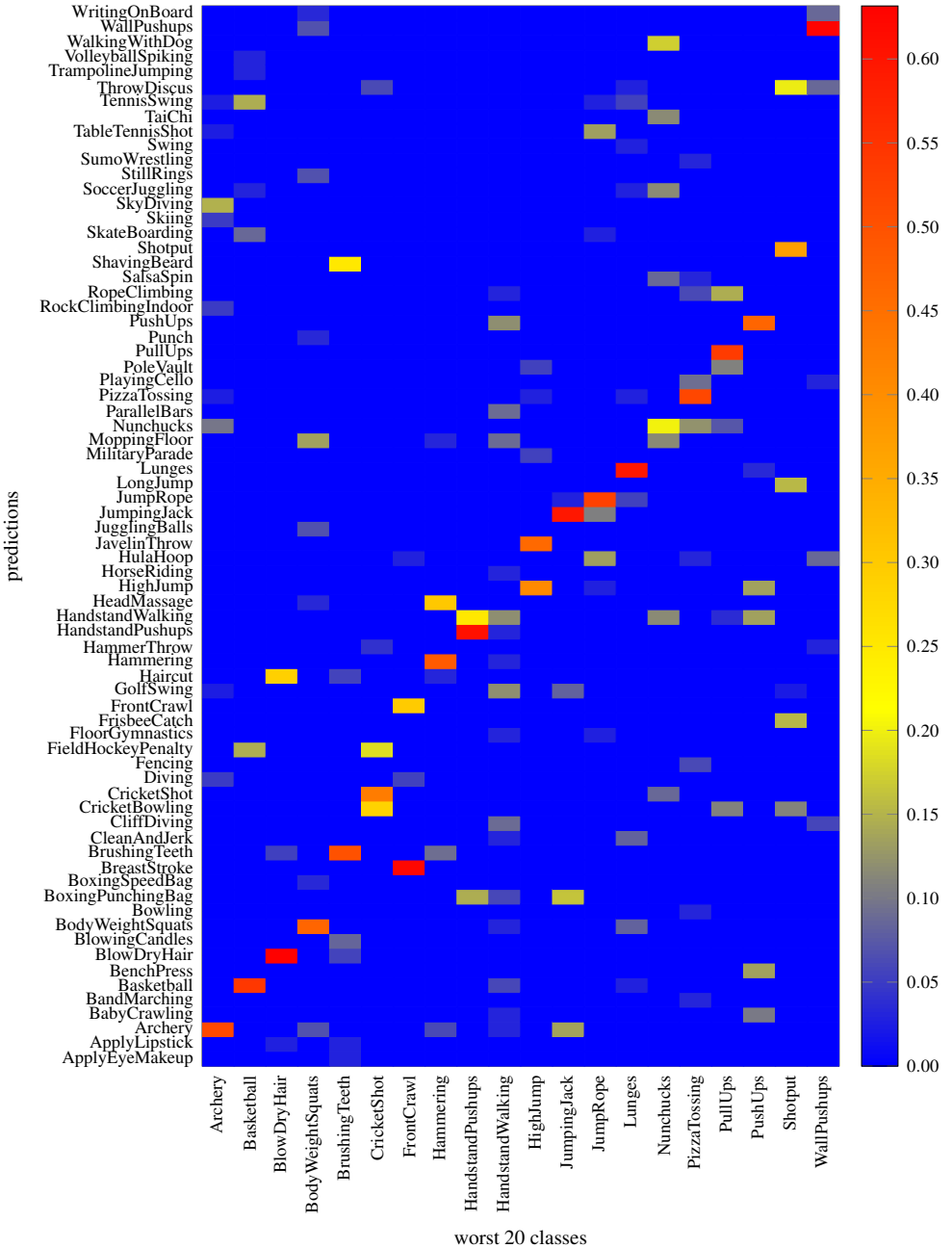


Figure 4: Confusions of the 20 worst performing action classes in spatial stream using "scene", "person" and "object" cues model.

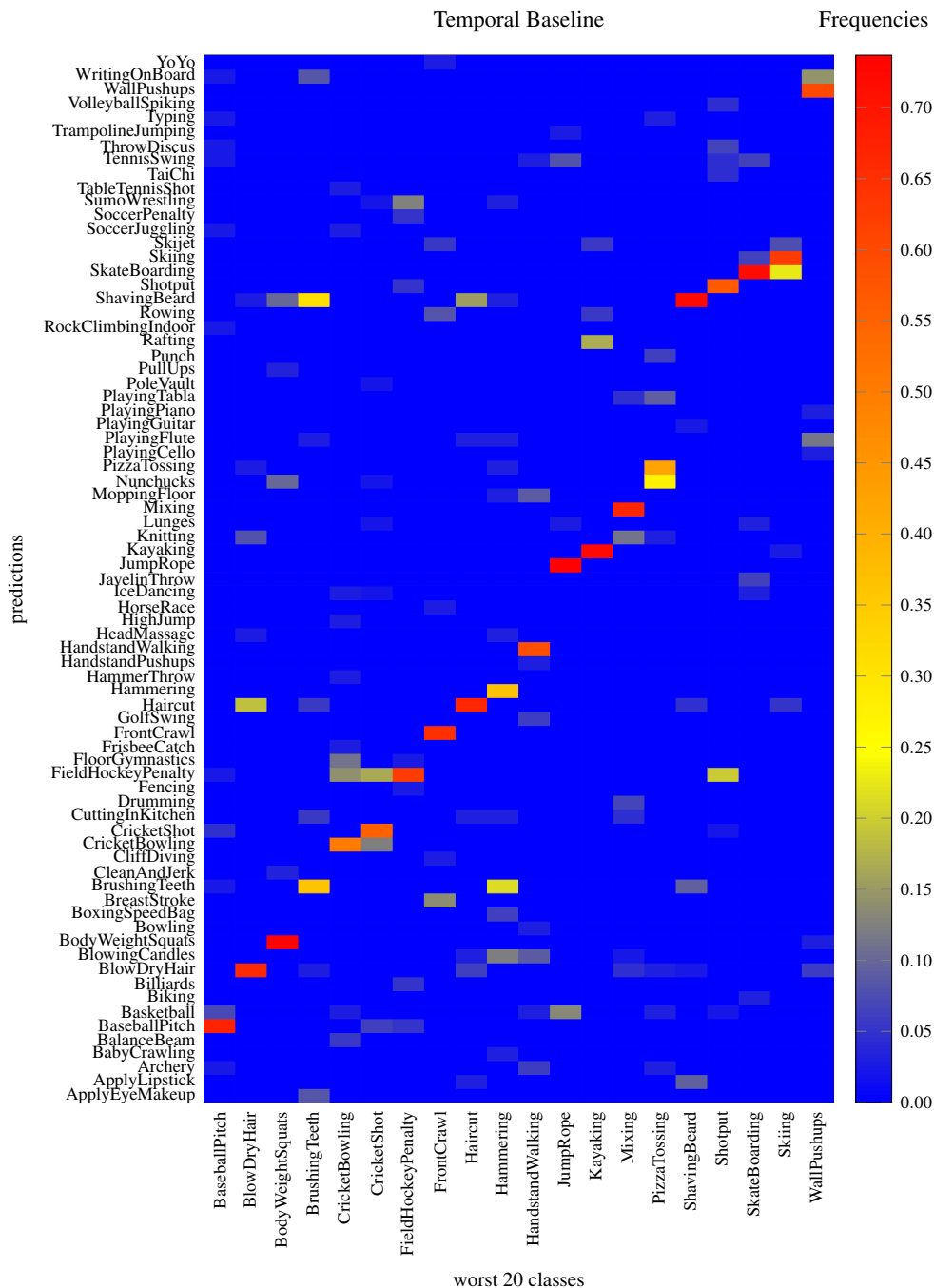
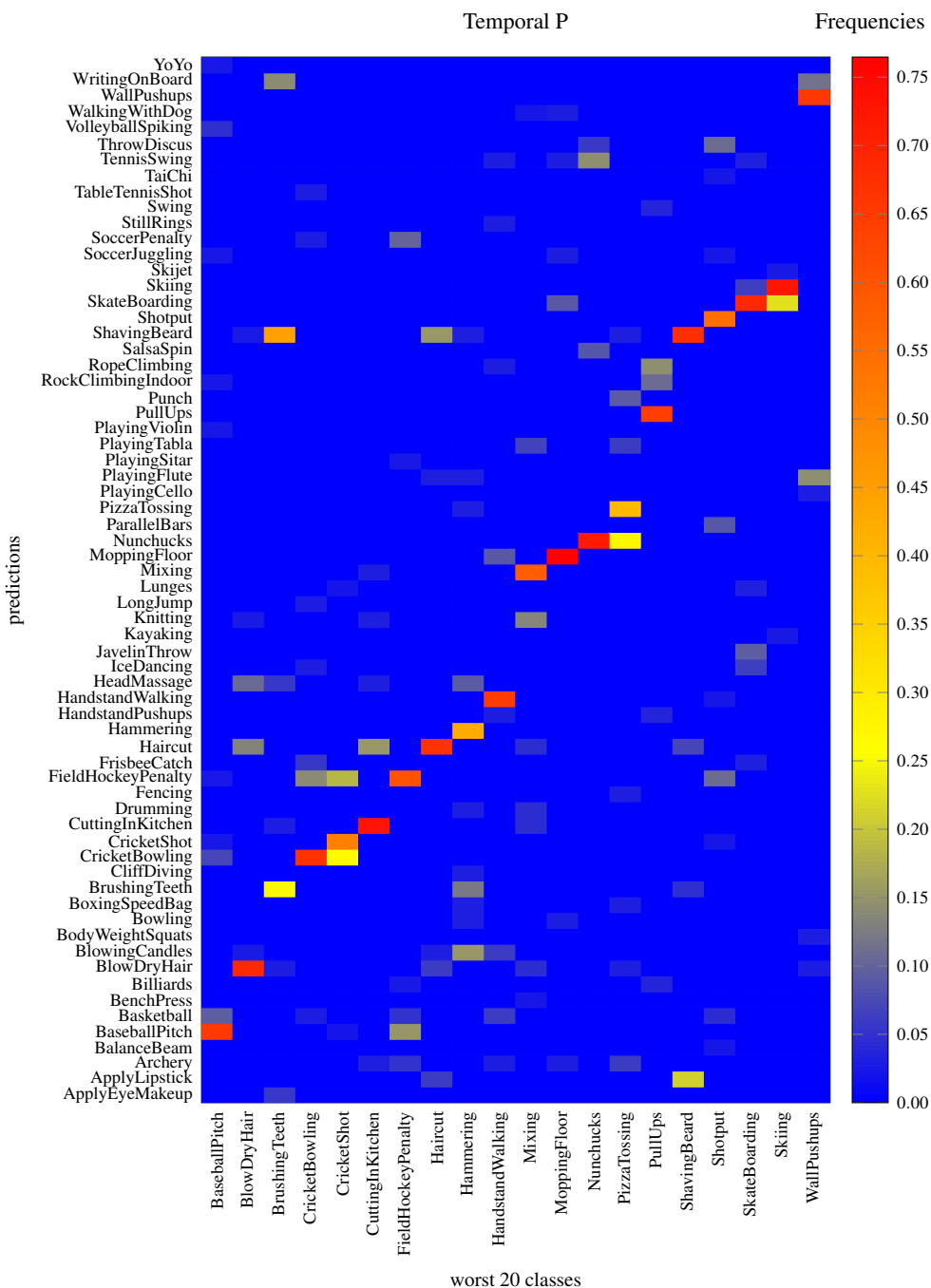


Figure 5: Confusions of the 20 worst performing action classes in temporal stream using baseline ("scene" only) model.





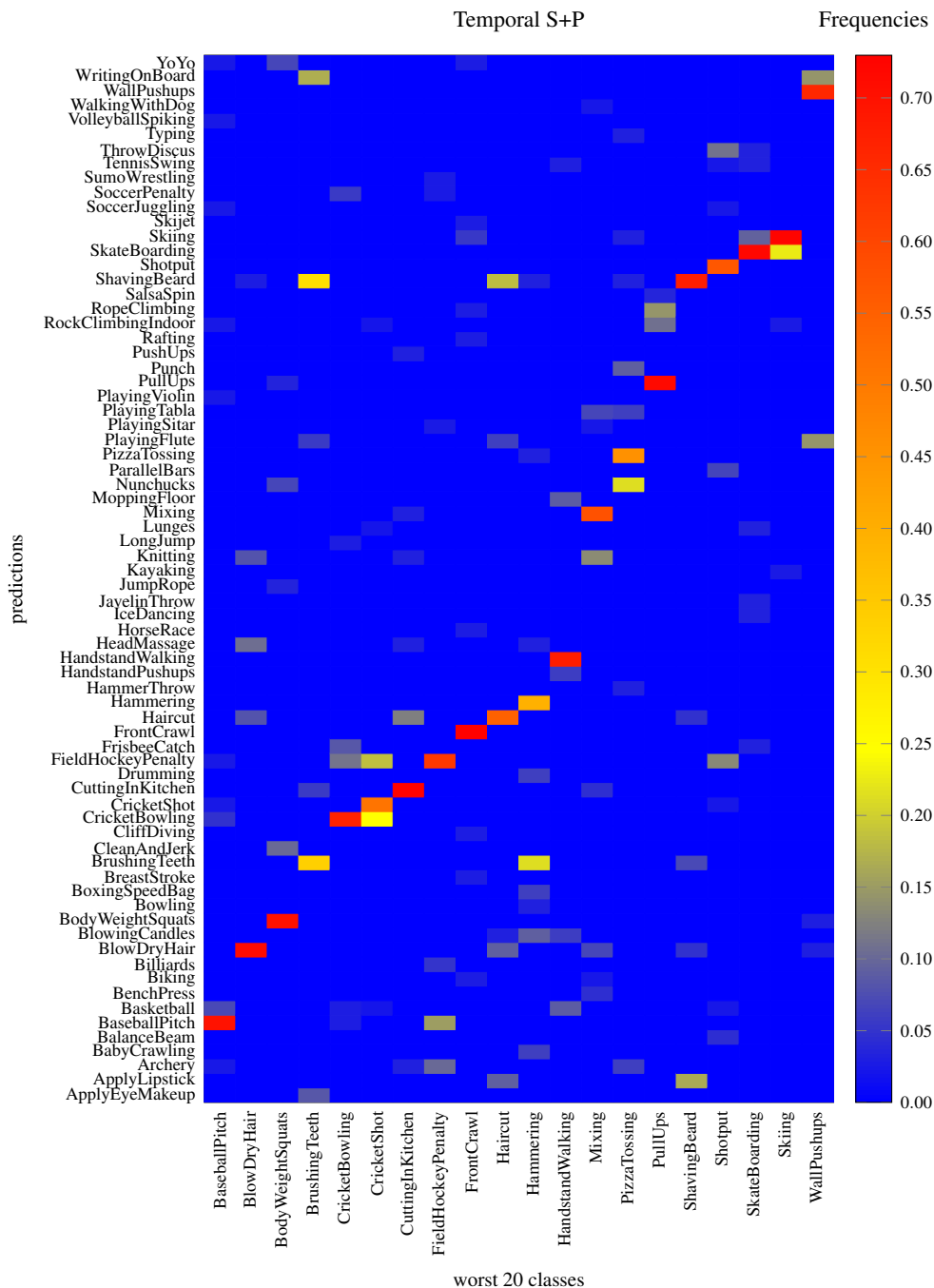


Figure 7: Confusions of the 20 worst performing action classes in temporal stream using "scene" and "person" cue (S+P model).