Context Aware Online Adaptation of Activity Recognition Models

Supplementary Materials

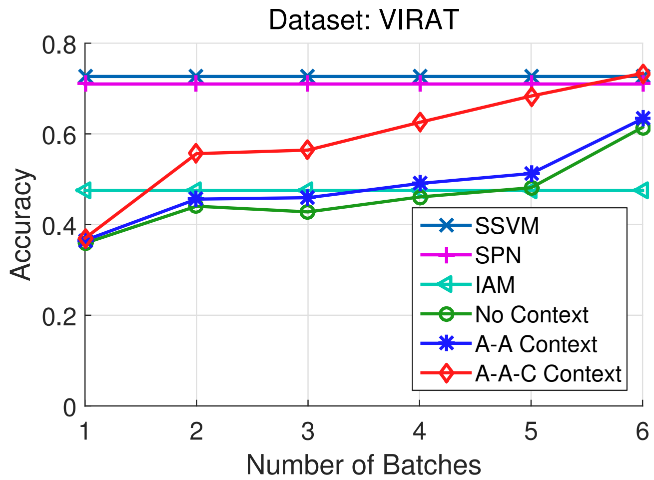
Anonymous ICCV submission

Paper ID 1873

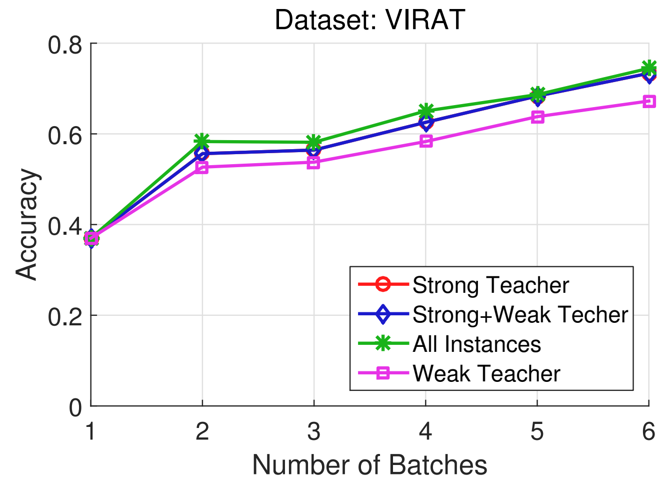
Table of Contents

1. High Resolution Plots

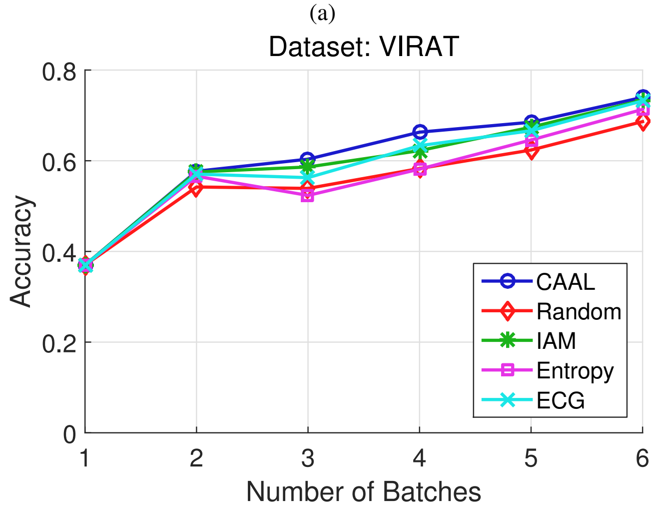
1.1. VIRAT



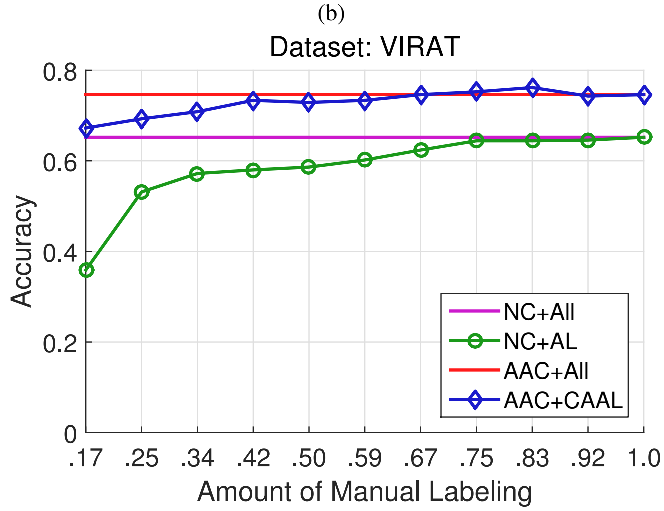
(a)



(b)



(c)



(d)

Figure 1. Plots a, b, c, and d show the performance evaluations for the VIRAT dataset. Plot (a) compares state-of-the-art batch and incremental methods against the three variants of the proposed approach. These variants are No Context (it only uses the appearance features with a softmax classifier), A-A context (it uses the spatial-temporal inter-relationship contexts of the activities represented by a CRF), and A-A-C context (it uses both of the associated object contexts and spatial-temporal inter-relationship context represented by a CRF). A-A-C test case performs better than all other methods with very less amount of manually labeled data. Plot (b) analyzes the performances of the four variants of our proposed active learning system. These variants are based on the use of two types of teachers. These variants are Strong Teacher (it manually labels a fraction of the incoming instances), Weak Teacher (it does not use any manually labeled data, only uses the highly confident labels provided by the classifier), Strong+Weak Teacher (it uses the labels provided by both of the above teachers.), and All Instances (it is the most expensive case, where all of the instances are manually labeled). Performance of the Strong+Weak Teacher is almost similar to All Instances but with very less amount of manually labeled data. It proves the robustness of our proposed framework. Plot (c) compares our proposed context aware active learning system (CAAL) against the recent active learning methods and random sampling. CAAL uses the labels provided by both of the strong and the weak teachers. Recent active learning methods are - incremental activity modeling (IAM), uncertainty of the nodes (Entropy), and expected change of gradients (ECG). Our proposed active learning system performs better than other active learning approaches. Plot (d) compares the accuracy against the amount of manual labeling. NC does not use any context information, whereas AAC use both of the spatial-temporal and the object contexts. NC+All and AAC+All test cases use all manually labeled data. It is evident that AAC+CAAL uses approximately fifty percent manually labeled data in order to achieve similar performance comparing to batch method that use all manually labeled data.

Go to Table of Contents

1.2. UCLA-Office

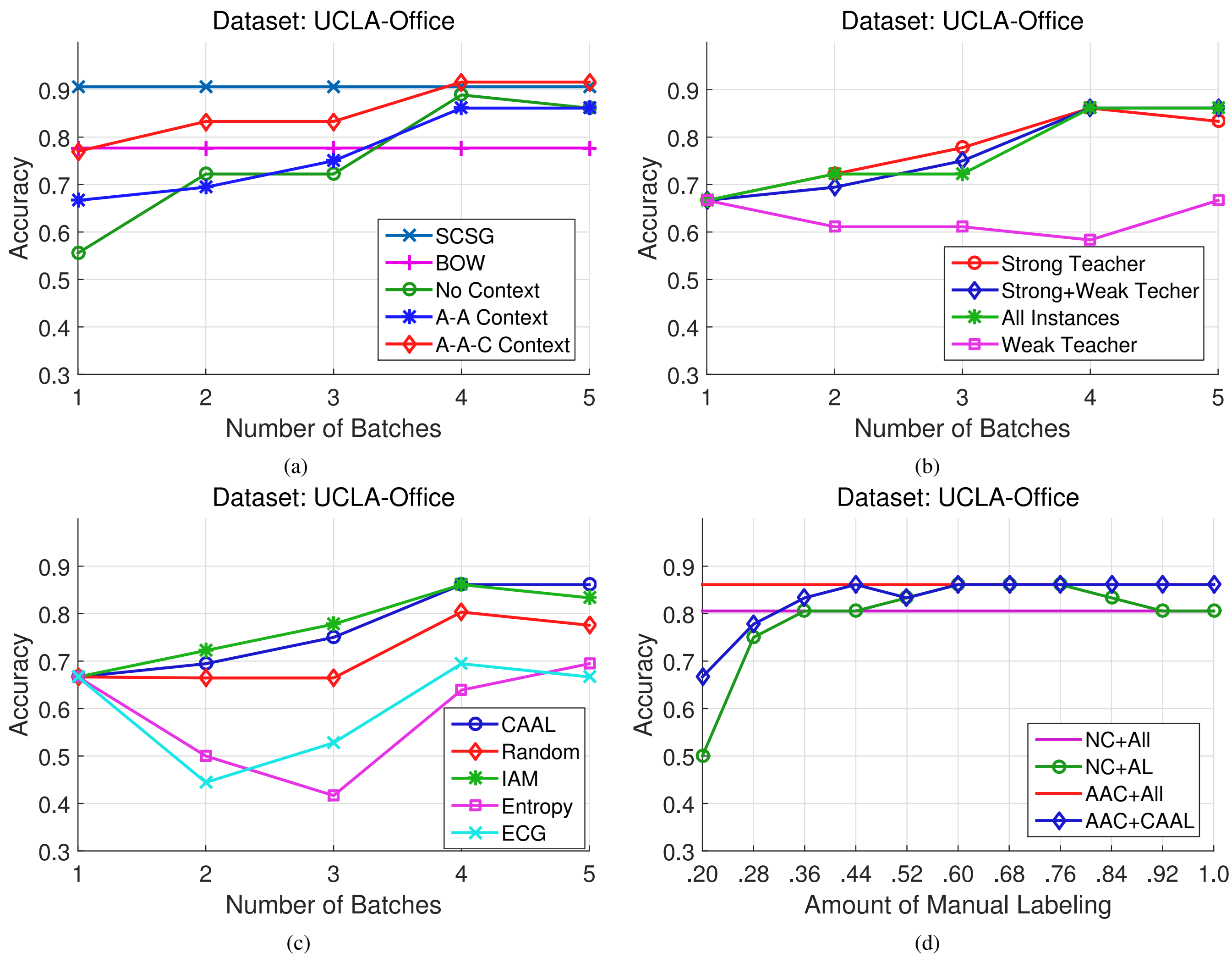


Figure 2. Plots a, b, c, and d show the performance evaluations for the UCLA-Office dataset. Plot (a) compares state-of-the-art batch and incremental methods against the three variants of the proposed approach. These variants are No Context (it only uses the appearance features with a softmax classifier), A-A context (it uses the spatial-temporal inter-relationship contexts of the activities represented by a CRF), and A-A-C context (it uses both of the associated object contexts and spatial-temporal inter-relationship context represented by a CRF). A-A-C test case performs better than all other methods with very less amount of manually labeled data. Plot (b) analyzes the performances of the four variants of our proposed active learning system. These variants are based on the use of two types of teachers. These variants are Strong Teacher (it manually labels a fraction of the incoming instances), Weak Teacher (it does not use any manually labeled data, only uses the highly confident labels provided by the classifier), Strong+Weak Teacher (it uses the labels provided by both of the above teachers.), and All Instances (it is the most expensive case, where all of the instances are manually labeled). Performance of the Strong+Weak Teacher is almost similar to All Instances but with very less amount of manually labeled data. It proves the robustness of our proposed framework. Plot (c) compares our proposed context aware active learning system (CAAL) against the recent active learning methods and random sampling. CAAL uses the labels provided by both of the strong and the weak teachers. Recent active learning methods are - incremental activity modeling (IAM), uncertainty of the nodes (Entropy), and expected change of gradients (ECG). Our proposed active learning system performs better than other active learning approaches. Plot (d) compares the accuracy against the amount of manual labeling. NC does not use any context information, whereas AAC use both of the spatial-temporal and the object contexts. NC+All and AAC+All test cases use all manually labeled data. It is evident that AAC+CAAL uses approximately forty percent manually labeled data in order to achieve similar performance comparing to batch method that use all manually labeled data.

Go to Table of Contents

ICCV #1873

1.3. MPII-Cooking

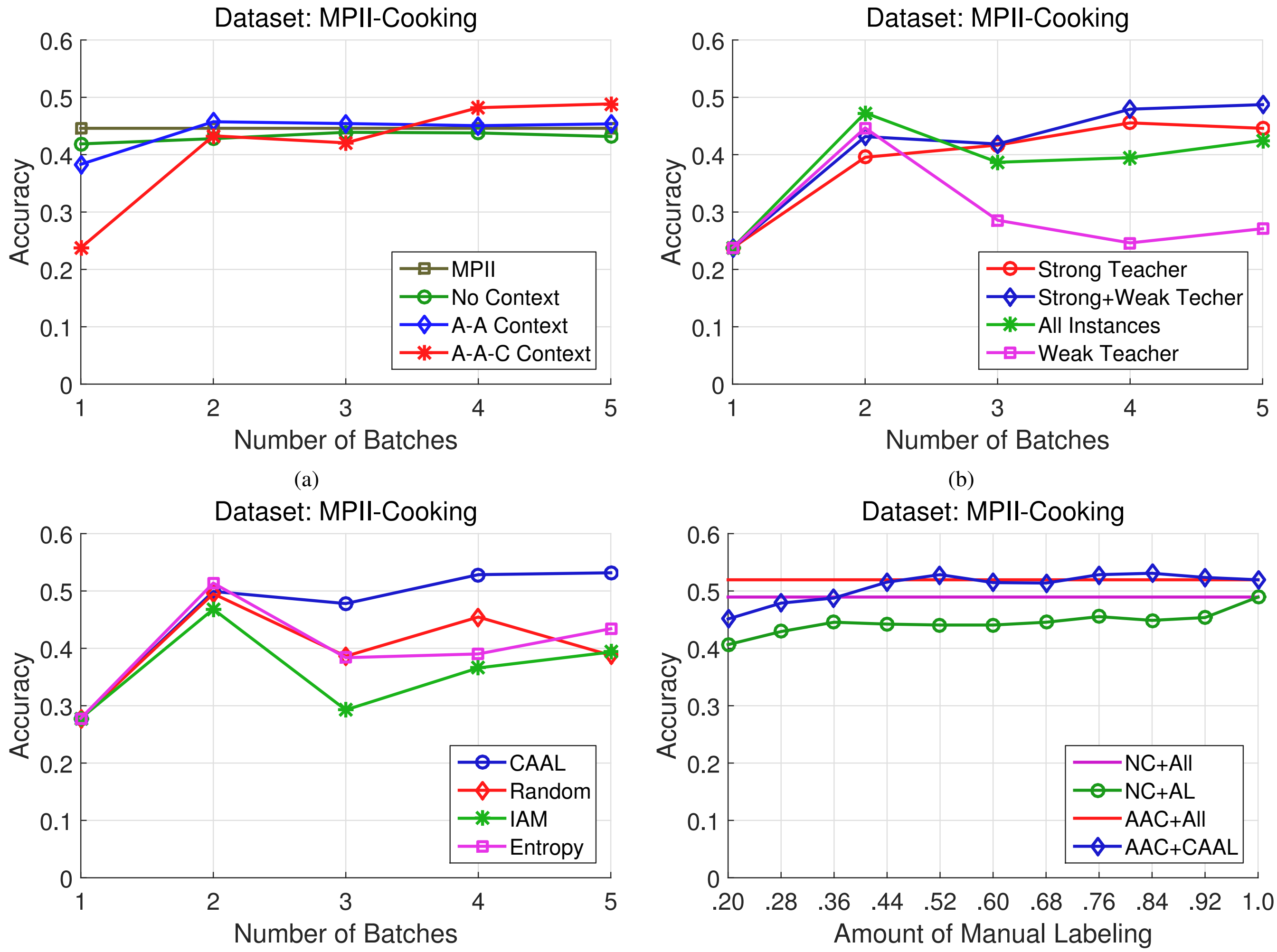
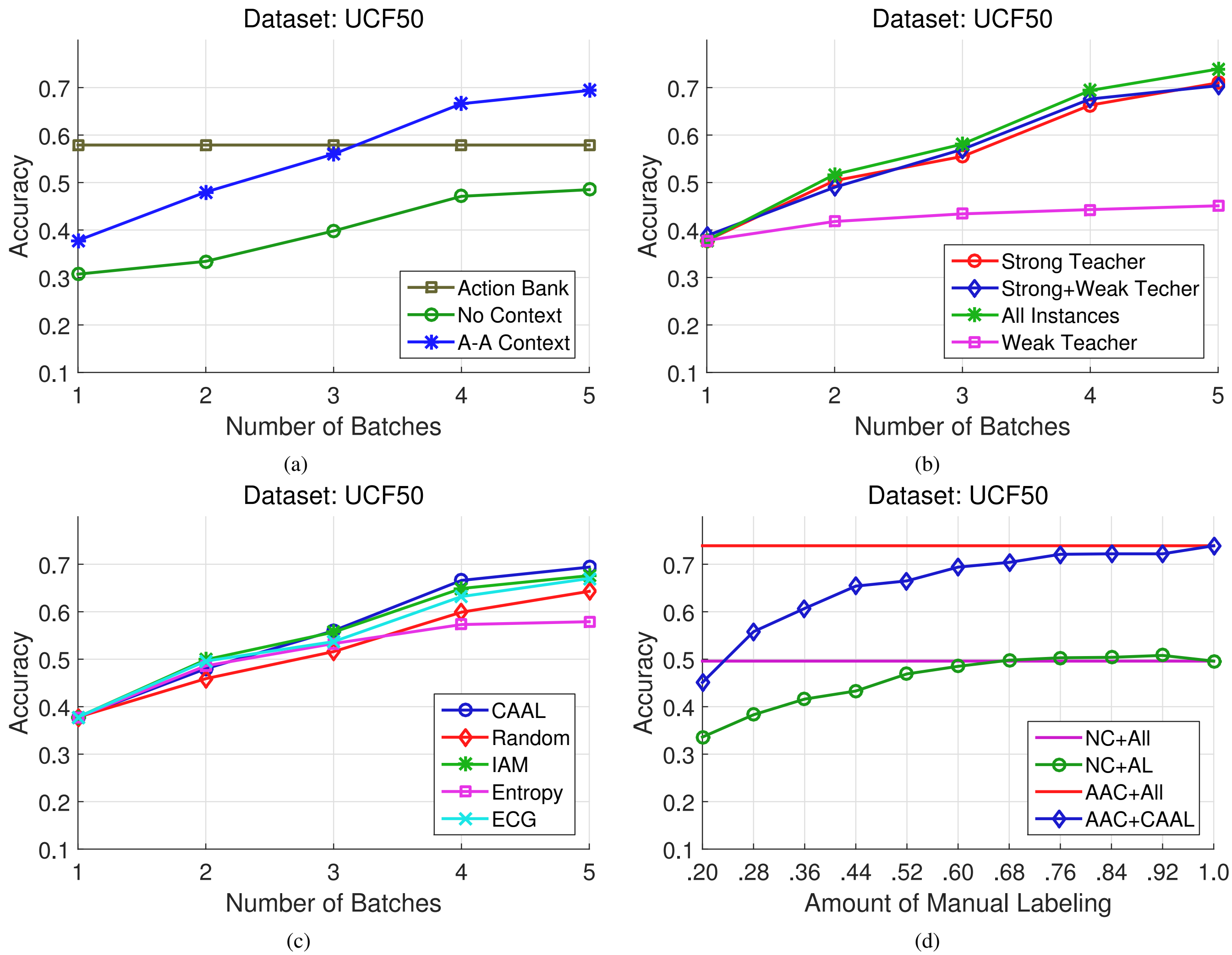


Figure 3. Plots a, b, c, and d show the performance evaluations for the MPII-Cooking dataset. Plot (a) compares state-of-the-art batch and incremental methods against the three variants of the proposed approach. These variants are No Context (it only uses the appearance features with a svm classifier), A-A context (it uses the spatial-temporal inter-relationship contexts of the activities represented by a CRF), and A-A-C context (it uses both of the associated object contexts and spatial-temporal inter-relationship context represented by a CRF). A-A-C test case performs better than all other methods with very less amount of manually labeled data. Plot (b) analyzes the performances of the four variants of our proposed active learning system. These variants are based on the use of two types of teachers. These variants are Strong Teacher (it manually labels a fraction of the incoming instances), Weak Teacher (it does not use any manually labeled data, only uses the highly confident labels provided by the classifier), Strong+Weak Teacher (it uses the labels provided by both of the above teachers.), and All Instances (it is the most expensive case, where all of the instances are manually labeled). Performance of the Strong+Weak Teacher is almost similar to All Instances but with very less amount of manually labeled data. It proves the robustness of our proposed framework. Plot (c) compares our proposed context aware active learning system (CAAL) against the recent active learning methods and random sampling. CAAL uses the labels provided by both of the strong and the weak teachers. Recent active learning methods are - incremental activity modeling (IAM), uncertainty of the nodes (Entropy), and expected change of gradients (ECG). Our proposed active learning system performs better than other active learning approaches. Plot (d) compares the accuracy against the amount of manual labeling. NC does not use any context information, whereas AAC use both of the spatial-temporal and the object contexts. NC+All and AAC+All test cases use all manually labeled data. It is evident that AAC+CAAL uses approximately forty percent manually labeled data in order to achieve similar performance comparing to batch method that use all manually labeled data.

Go to Table of Contents

1.4. UCF50



(a) (b) (c) (d)

Figure 4. Plots a, b, c, and d show the performance evaluations for the UCF50 dataset. Plot (a) compares state-of-the-art batch and incremental methods against the three variants of the proposed approach. These variants are No Context (it only uses the appearance features with a softmax classifier) and A-A context (it uses the spatial-temporal inter-relationship contexts of the activities represented by a CRF). A-A test case performs better than all other methods with very less amount of manually labeled data. Plot (b) analyzes the performances of the four variants of our proposed active learning system. These variants are based on the use of two types of teachers. These variants are Strong Teacher (it manually labels a fraction of the incoming instances), Weak Teacher (it does not use any manually labeled data, only uses the highly confident labels provided by the classifier), Strong+Weak Teacher (it uses the labels provided by both of the above teachers.), and All Instances (it is the most expensive case, where all of the instances are manually labeled). Performance of the Strong+Weak Teacher is almost similar to All Instances but with very less amount of manually labeled data. It proves the robustness of our proposed framework. Plot (c) compares our proposed context aware active learning system (CAAL) against the recent active learning methods and random sampling. CAAL uses the labels provided by both of the strong and the weak teachers. Recent active learning methods are - incremental activity modeling (IAM), uncertainty of the nodes (Entropy), and expected change of gradients (ECG). Our proposed active learning system performs better than other active learning approaches. Plot (d) compares the accuracy against the amount of manual labeling. NC does not use any context information, whereas AAC use both of the spatial-temporal and the object contexts. NC+All and AAC+All test cases use all manually labeled data.

Go to Table of Contents

2. Dataset Description

2.1. VIRAT

Description	The VIRAT dataset is a state-of-the-art human action dataset with many challenging characteristics such as wide variation in the activities and a high amount of occlusion and clutter. It consists of surveillance videos such as parking lot videos involving single vehicle activities, person and vehicle interactions, and people interactions. There are also some group activities. This dataset consists of scenes captured on a single camera although the viewpoint can differ from one scene to the next. In any scene, the activities can occur at different orientations depending on the location. However, since these are wide-area videos, persons of interest are usually far away from the camera.
Number of Scenes	11
Number of Sequences	329
Number of Activities	1555
Number of Activity Types	11
Activity Types	Loading an object to a vehicle, Unloading an object from a vehicle, Opening a vehicle trunk, Closing a vehicle trunk, Getting into a vehicle, Getting out of a vehicle, Gesturing, Person carrying an object, Person running, Person entering a facility, and Person exiting a facility.
Associated Object Types	Person, Car, Vehicle, Carrying objects, and Bike.
Video Resolution	1920×1080
Total Video Duration	About 5 hours
Wild?	Yes
Segmented?	No
Background	Fixed for a sequence

Go to Table of Contents

2.2. UCLA-Office

Description	The UCLA office dataset consists of indoor and outdoor videos of single and two-person activities. Here, we perform experiments on the lab scene containing close to 35 minutes of video captured with a single fixed camera in a room. There is very little variation in viewpoint, occlusion and scale here. Each activity occurs 6 to 15 times in the dataset.
Number of Scenes	1
Number of Sequences	3
Number of Activities	157
Number of Activity Types	10
Activity Types	EnterRoom, ExitRoom, SitDown, StandUp, WorkLaptop, WorkPaper, ThrowTrash, PourDrink, PickPhone, PlacePhone.
Associated Object Types	Laptop, Phone, Paper, Trash, etc.
Video Resolution	1280×720
Total Video Duration	35 minutes
Wild?	No
Segmented?	No
Background	Fixed

Go to Table of Contents

2.3. MPII-Cooking

Description	This dataset contains fine-grained cooking activities in an indoor settings. 12 participants performed 65 different cooking activities. Participants were asked to prepare one to six of a total of 14 dishes such as fruit salad or cake containing several cooking activities. Activities have low inter-class variability and high intra-class variability due to diverse subjects and cooking ingredients. However, activities have very little occlusion, clutter, or change of viewpoints.
Number of Subjects	12
Number of Sequences	44
Number of Activities	5609
Number of Activity Types	65
Activity Types	background, changeTemp, cutApart, cutDice, cutIn, cutOffEnds, cutOutInside, cutSlices, cutStripes, dry, fillWaterFromTap, grate, lidPutOn, lidRemove, mix, move, openEgg, openTin, openCloseCupboard, openCloseDrawer, openCloseFridge, openCloseOven, package, peel, plugInOut, pour, pullOut, puree, putInBowl, putInPanPot, putOnBreadDough, putOnCuttingBoard, putOnPlate, read, removeFromPackage, ripOpen, scratchOff, screwClose, screwOpen, shake, smell, spice, spread, squeeze, stamp, stir, strew, takePutInCupboard, takePutInDrawer, takePutInFridge, takePutInOven, TakePutInSpiceHolder, takeIngredientApart, takeOutFromCupboard, takeOutFromDrawer, takeOutFromFridge, takeOutFromOven, takeOutFromSpiceHolder, taste, throwInGarbage, unrollDough, washHands, washObjects, whisk, wipeClean.
Associated Object Types	hand, bottle, cap opener, mug, beer, tomato, etc.
Video Resolution	1624×1224
Total Video Duration	8 hours 20 minutes
Wild?	No
Segmented?	No
Background	Fixed

Go to Table of Contents

2.4. UCF50

Description	This dataset contains various types of sports activities in the wild. It has very low inter-class and very high intra-class variability. Activities have very high amount of occlusion, clutter and change of viewpoints.
Number of Subjects	25
Number of Sequences	6676
Number of Activities	6676
Number of Activity Types	50
Activity Types	BaseballPitch, Basketball, BenchPress, Biking, Billiards, BreastStroke, CleanAndJerk, Diving, Drumming, Fencing, GolfSwing, HighJump, HorseRace, HorseRiding, HulaHoop, JavelinThrow, JugglingBalls, JumpRope, JumpingJack, Kayaking, Lunges, MilitaryParade, Mixing, Nunchucks, PizzaTossing, PlayingGuitar, PlayingPiano, PlayingTabla, PlayingViolin, PoleVault, PommelHorse, PullUps, Punch, PushUps, RockClimbingIndoor, RopeClimbing, Rowing, SalsaSpin, SkateBoarding, Skiing, Skijet, SoccerJuggling, Swing, TaiChi, TennisSwing, ThrowDiscus, TrampolineJumping, VolleyballSpiking, WalkingWithDog, YoYo.
Associated Object Types	—
Video Resolution	320×240
Total Video Duration	About 8 hours
Wild?	Yes
Segmented?	Yes
Background	Dynamic

Go to Table of Contents

3. Context Features

Contexts may vary based on the application domain. Our proposed generalized framework can handle any number and types of contexts. In this paper, we use different contexts for different datasets. We describe them below.

3.1. VIRAT

We use both of the spatial-temporal relationships among the activities context and object contexts for this dataset. These contextual features have been described in the main paper in Equations 2-9.

3.2. UCLA-Office

We only use the spatial-temporal relationships among the activities context for this dataset. This context feature has been described in the main paper in Equations 2 and 6.

3.3. MPII-Cooking

We use both of the spatial-temporal relationships among the activities context and object contexts for this dataset. Spatial-temporal relationship context remains the same as in Equations 2 and 6 in the main paper. Activities in this dataset involve three types of objects - tools (c_i^1), ingredients (c_i^2), and containers (c_i^3). We use each of them as a separate context and formulate them as like in Equations 3, 4, 7, and 8. So the Equations 3 and 7 will become,

$$\phi(c_i, z_i) = \phi(c_i^1, z_i) \odot \phi(c_i^2, z_i) \odot \phi(c_i^3, z_i)$$
$$\psi(a_i, c_i) = \psi(a_i, c_i^1) \otimes \psi(a_i, c_i^2) \otimes \psi(a_i, c_i^3)$$

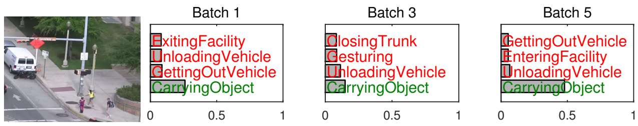
3.4. UCF50

Since the activities in UCF50 dataset are segmented, there are no natural spatial-temporal relationships exist among the activities. Also, each activity involves a person and a particular tool. Use of object context might overfit the model. So, we improvise a relationship among the activities. We roughly categorize fifty activity classes into eight groups and assume that in each group activities are inter-related. These groups are - Outdoor Group Sports (BaseballPitch, Basketball, VollyballSpiking, TennisSwing, HorseRace, and Rowing), Outdoor Individual Sports (GolfSwing, HighJump, JavelinThrow, Kayaking, Skiing, SoccerJuggling, ThrowDiscuss, and PoleVolt), Indoor Sports (Billiards, CleanAndJerk, Fencing, PommelHorse, Punch, and RockClimbing), Outdoor Activity (Biking, Diving, MilitaryParade, NunChucks, HorseRiding, RopeClimbing, SkateBoarding, SkiJet, Swing, and TampolineJumping), Indoor Activity (SalsaSpin, BreastStroke, HulaHoop, JugglingBalls, and YoYo), Physical Exercise (BenchPress, JumpingJack, JumpRope, TaiChi, Walking, PullUps, PushUps, and Lunges), Kitchen (Mixing and PizzaTossing), and Instrumental (Drumming, PlayingGuitar, PlayingPiano, PlayingTabla, and PlayingViolin). The corresponding mathematical formulations remain same as like Equations 2 and 6 of the main paper.

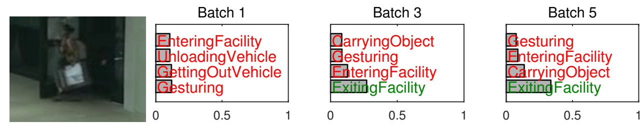
Go to Table of Contents

4. Evaluation of Continuous Learning on Individual Activities

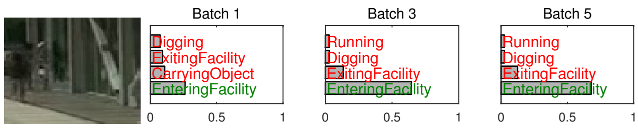
4.1. VIRAT



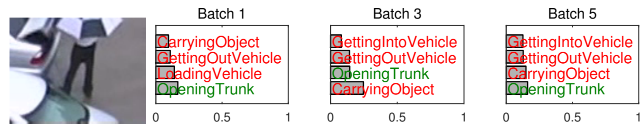
(a) Person Carrying an Object



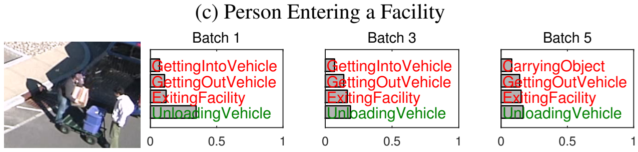
(b) Person Exiting a Facility



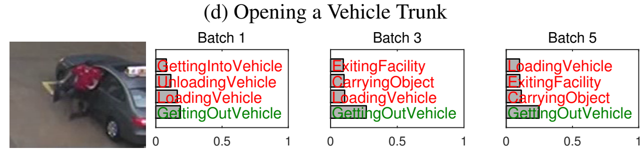
(c) Person Entering a Facility



(d) Opening a Vehicle Trunk

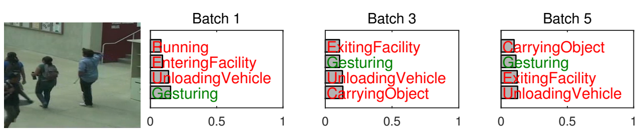


(e) Unloading an Object from a Vehicle

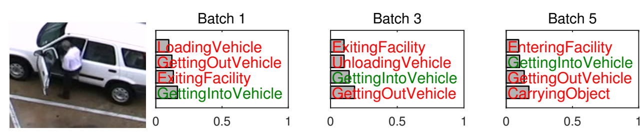


(f) Getting Out of a Vehicle

Figure 5. Evaluation of continuous learning on individual activities. Activity with green color means the ground truth class, whereas activities with red color means false predictions. Grey bars represent probability scores. Here, we show the results obtained after the arrival of batch 1, 3, and 5 data. The plots in this figure show some of the successful examples, where continuous learning helps to obtain the correct label with a higher probability even though some of them were miss-classified initially. Best viewable in color.



(a) Person Gesturing

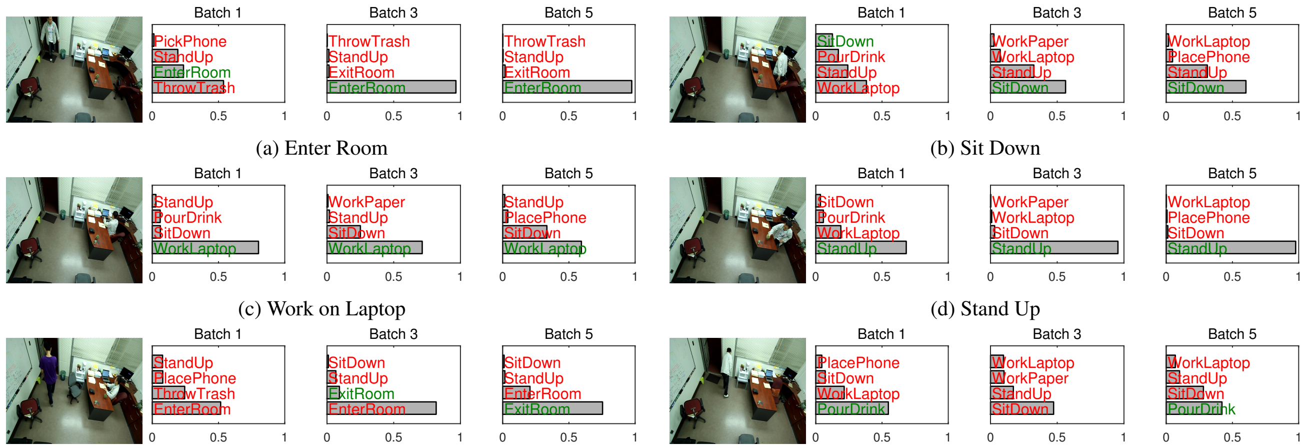


(b) Getting Into a Vehicle

Figure 6. Evaluation of continuous learning on individual activities. Activity with green color means the ground truth class, whereas activities with red color means false predictions. Grey bars represent probability scores. Here, we show the results obtained after the arrival of batch 1, 3, and 5 data. The plots in this figure show few failure cases. Best viewable in color.

Go to Table of Contents

4.2. UCLA-Office

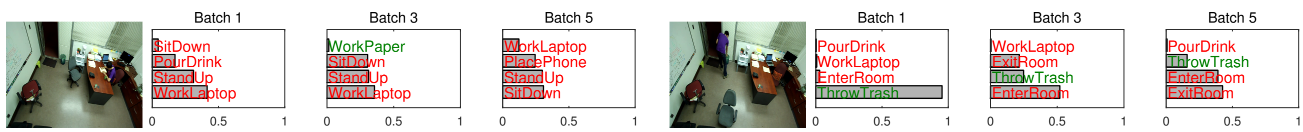


(a) Enter Room (b) Sit Down

(c) Work on Laptop (d) Stand Up

(e) Exit Room (f) Pour Drink

Figure 7. Evaluation of continuous learning on individual activities. Activity with green color means the ground truth class, whereas activities with red color means false predictions. Grey bars represent probability scores. Here, we show the results obtained after the arrival of batch 1, 3, and 5 data. The plots in this figure show some of the successful examples, where continuous learning helps to obtain the correct label with a higher probability even though some of them were miss-classified initially. Best viewable in color.

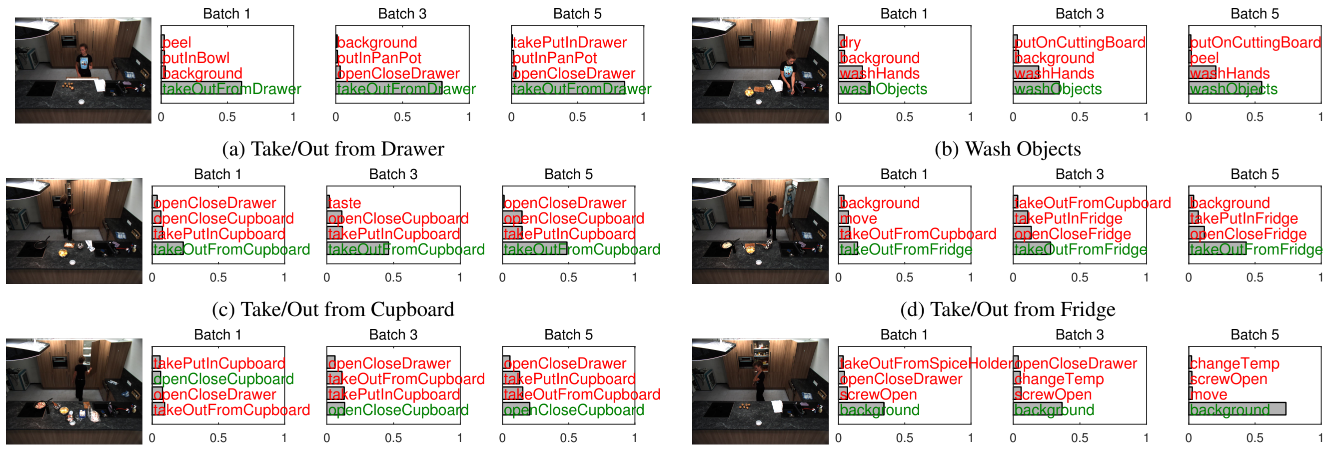


(a) Work on Paper (b) Throw Trash

Figure 8. Evaluation of continuous learning on individual activities. Activity with green color means the ground truth class, whereas activities with red color means false predictions. Grey bars represent probability scores. Here, we show the results obtained after the arrival of batch 1, 3, and 5 data. The plots in this figure show few failure cases. Best viewable in color.

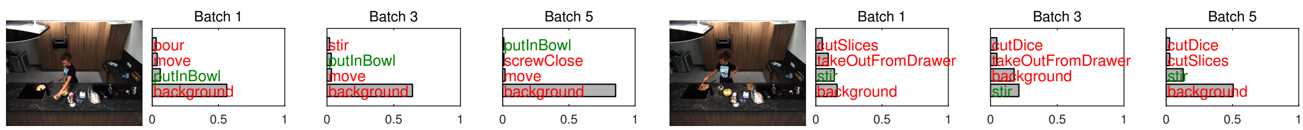
Go to Table of Contents

4.3. MPII-Cooking



(a) Take/Out from Drawer (b) Wash Objects
(c) Take/Out from Cupboard (d) Take/Out from Fridge
(e) Open/Close Cupboard (f) Background

Figure 9. Evaluation of continuous learning on individual activities. Activity with green color means the ground truth class, whereas activities with red color means false predictions. Grey bars represent probability scores. Here, we show the results obtained after the arrival of batch 1, 3, and 5 data. The plots in this figure show some of the successful examples, where continuous learning helps to obtain the correct label with a higher probability even though some of them were miss-classified initially. Best viewable in color.

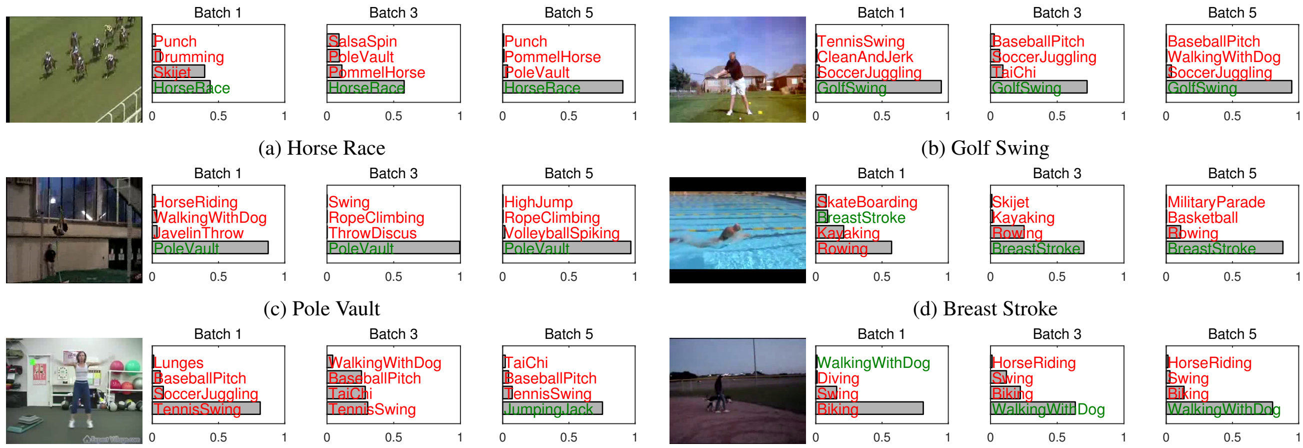


(a) Put in Bowl (b) Stir

Figure 10. Evaluation of continuous learning on individual activities. Activity with green color means the ground truth class, whereas activities with red color means false predictions. Grey bars represent probability scores. Here, we show the results obtained after the arrival of batch 1, 3, and 5 data. The plots in this figure show few failure cases. Best viewable in color.

Go to Table of Contents

4.4. UCF50

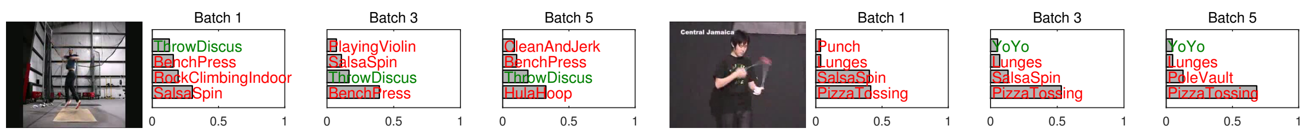


(a) Horse Race (b) Golf Swing

(c) Pole Vault (d) Breast Stroke

(e) Jumping Jack (f) Walking With Dog

Figure 11. Evaluation of continuous learning on individual activities. Activity with green color means the ground truth class, whereas activities with red color means false predictions. Grey bars represent probability scores. Here, we show the results obtained after the arrival of batch 1, 3, and 5 data. The plots in this figure show some of the successful examples, where continuous learning helps to obtain the correct label with a higher probability even though some of them were miss-classified initially. Best viewable in color.



(a) Throw Discus (b) Yo Yo

Figure 12. Evaluation of continuous learning on individual activities. Activity with green color means the ground truth class, whereas activities with red color means false predictions. Grey bars represent probability scores. Here, we show the results obtained after the arrival of batch 1, 3, and 5 data. The plots in this figure show few failure cases. Best viewable in color.

Go to Table of Contents

5. Activity Class-wise Performance Evaluation

5.1. VIRAT

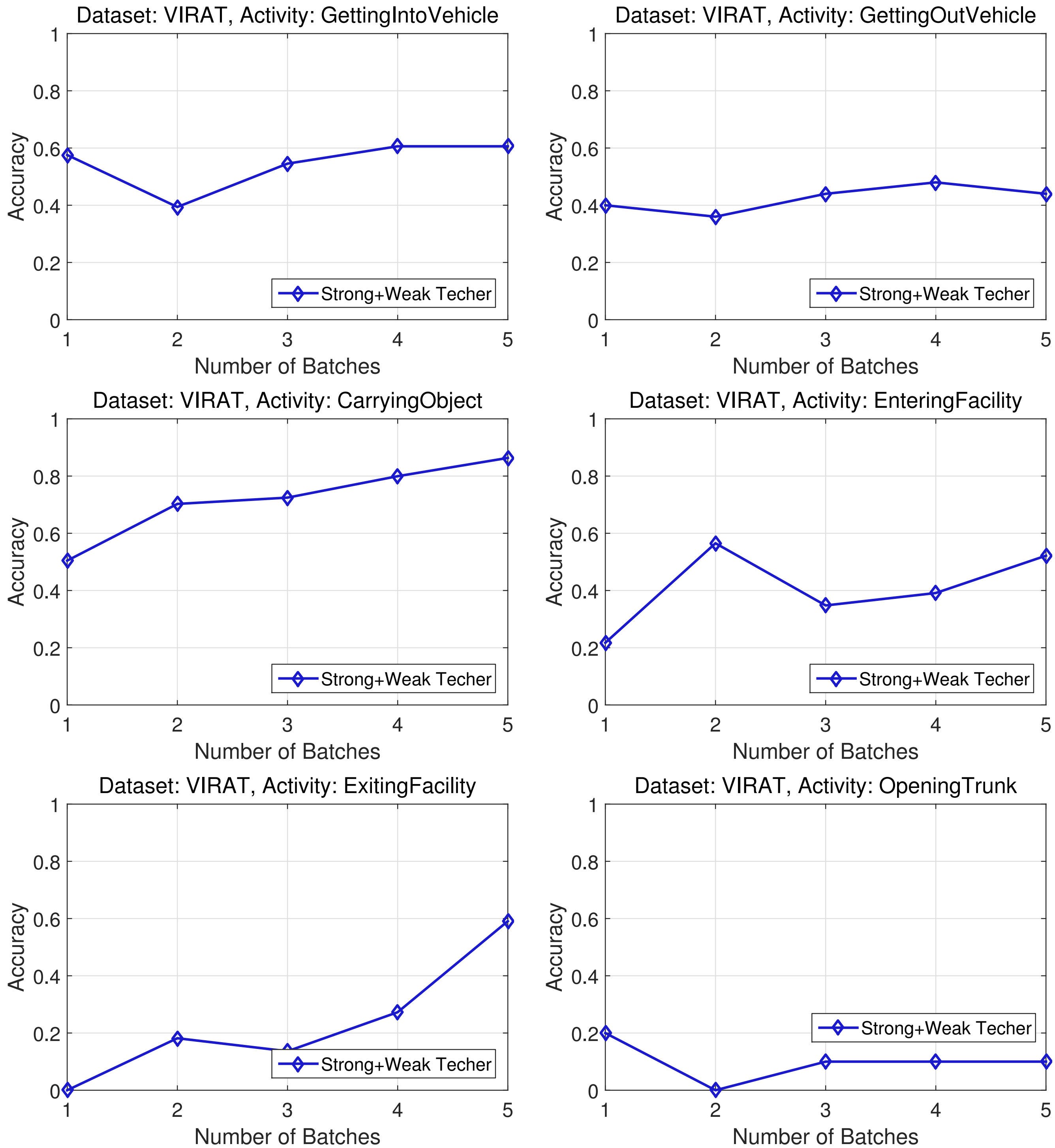


Figure 13. Activity class-wise performance evaluation of our proposed framework. This figure shows some selected activity classes of VIRAT dataset. In most of the cases, recognition performance of our framework for a particular class improves with the availability of new training data.

Go to Table of Contents

5.2. UCLA-Office

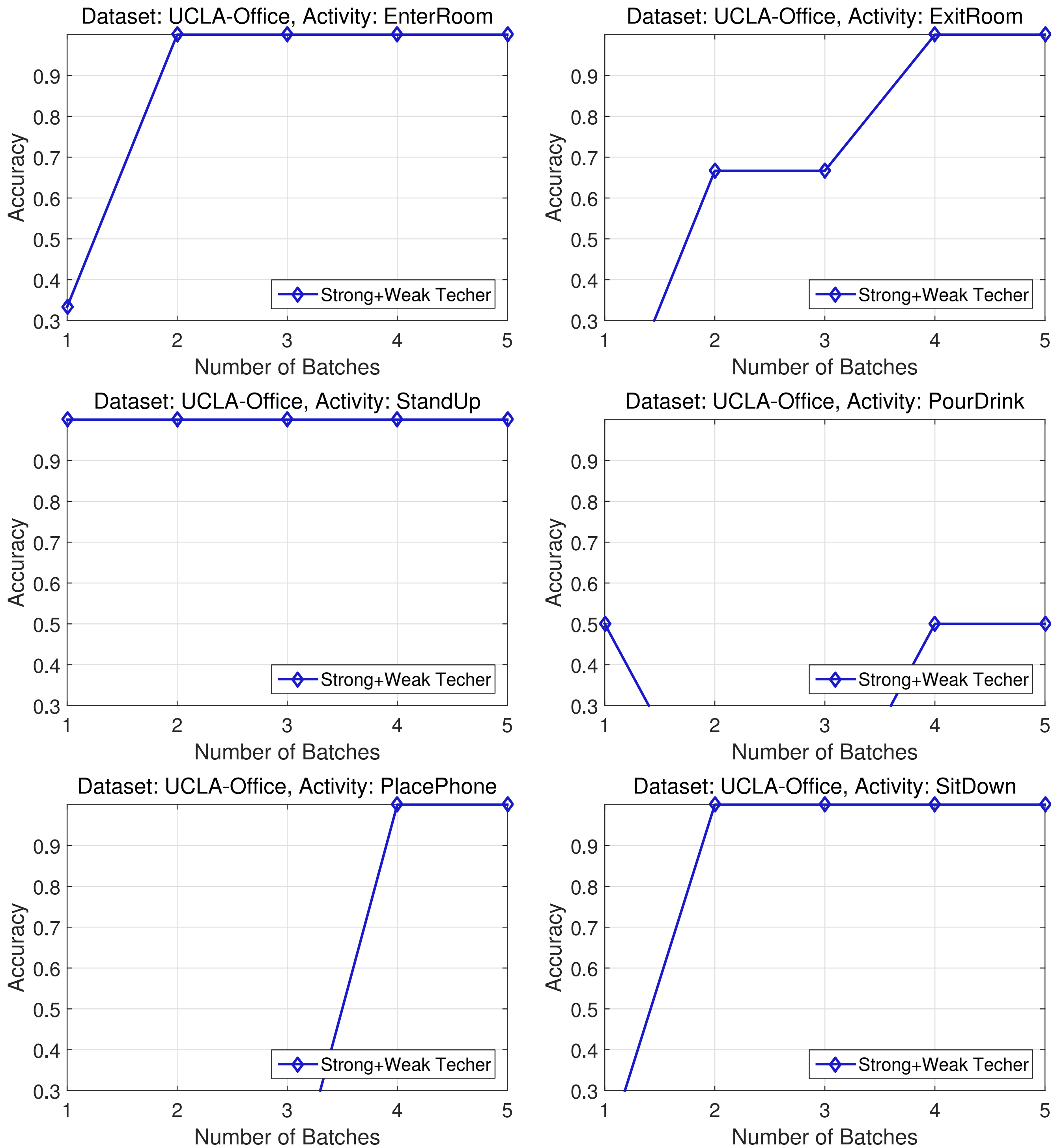


Figure 14. Activity class-wise performance evaluation of our proposed framework. This figure shows some selected activity classes of UCLA-Office dataset. In most of the cases, recognition performance of our framework for a particular class improves with the availability of new training data.

Go to Table of Contents

5.3. MPII-Cooking

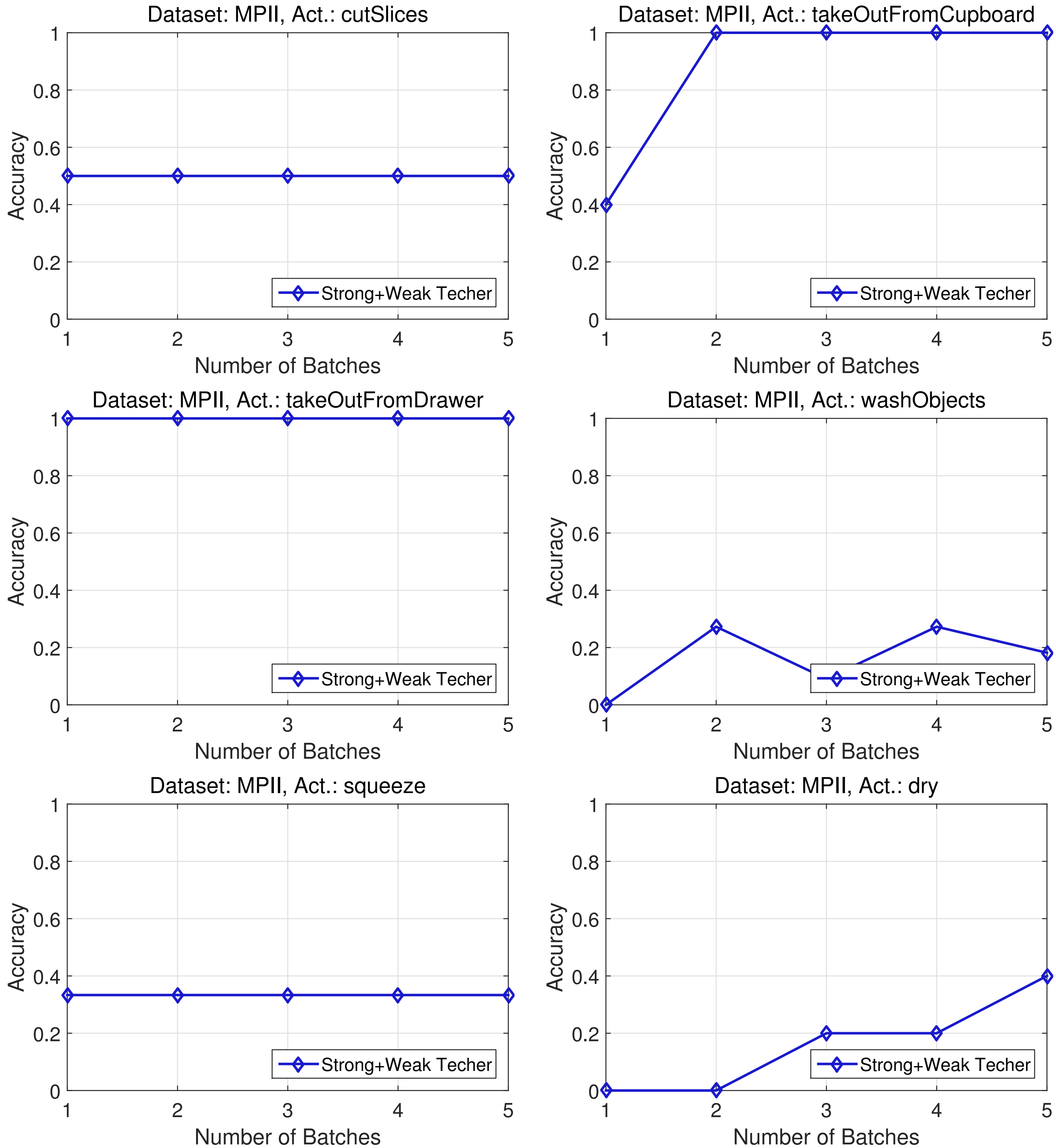


Figure 15. Activity class-wise performance evaluation of our proposed framework. This figure shows some selected activity classes of MPII-Cooking dataset. In most of the cases, recognition performance of our framework for a particular class improves with the availability of new training data.

Go to Table of Contents

5.4. UCF50

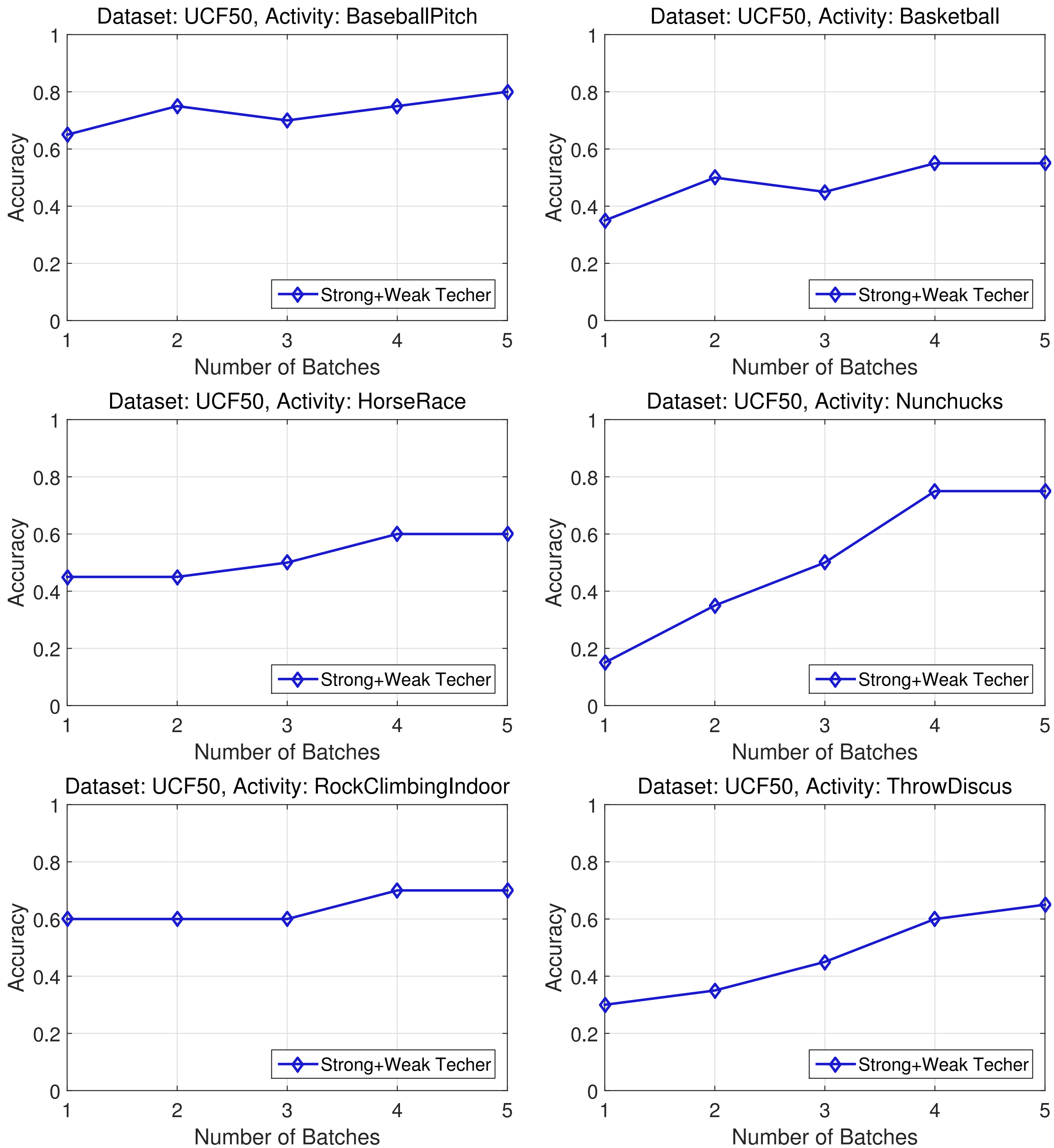


Figure 16. Activity class-wise performance evaluation of our proposed framework. This figure shows some selected activity classes of UCF50 dataset. In most of the cases, recognition performance of our framework for a particular class improves with the availability of new training data.

Go to Table of Contents

6. Parameter Values

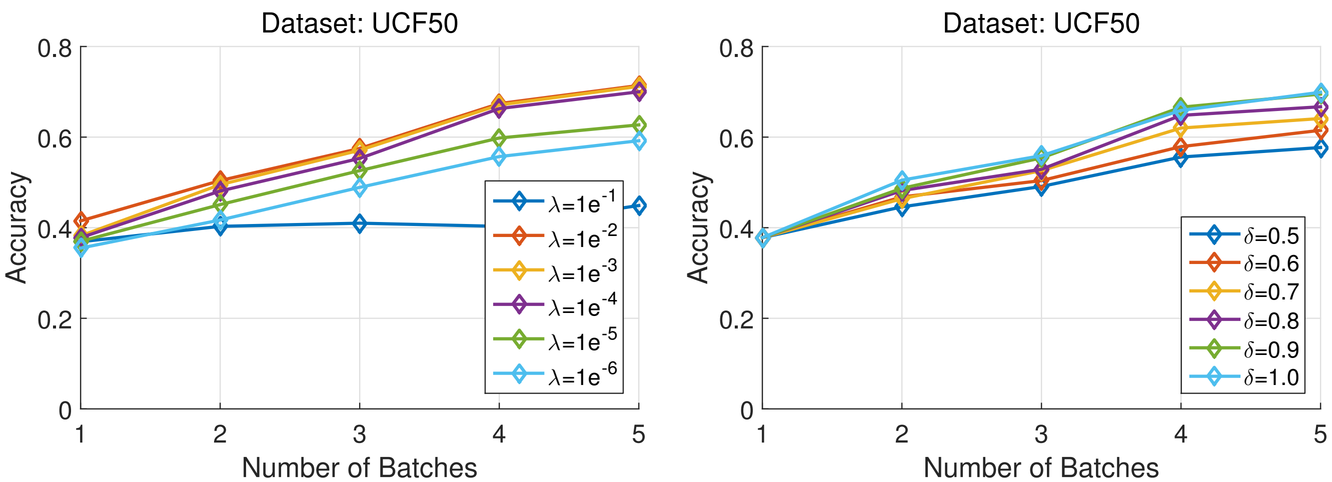
We learn most of the parameters from training data. We manually set only three parameters - manual labeling percentage (K), weight decay parameter (λ) of our baseline softmax classifier (Section 4.1 of the main paper), and the weak teacher threshold parameter (δ). We provided the sensitivity analysis of K as the accuracy vs. manual labeling plot in the Figure 5(d, h, l, and p) of the main paper for all datasets. Here, we present the values of K, λ, and δ that we used during our experiments for all datasets followed by the sensitivity analysis of λ and δ for VIRAT dataset.

	Dataset			
Parameters	VIRAT	UCLA-Office	MPII-Cooking	UCF50
K	0.5	0.5	0.5	0.5
λ	10^{-1}	10^{-2}	(used lin. svm)	10^{-4}
δ	0.9	0.9	0.9	0.9

Table 1. Parameter Values

6.1. Parameter Sensitivity

6.1.1 VIRAT



(a) (b)

Figure 17. Plot (a) illustrates the sensitivity analysis of the parameter λ. It shows that the choice of λ has significant effect on the performance, however performance of the framework is quite similar in range of $10^{-2} - 10^{-4}$. Plot (b) illustrates the sensitivity analysis of the parameter δ. Our framework performs better for the higher values of δ. It means that for a higher value of δ the framework will use very high confident labels from the classifier to retrain it. For a lower value of δ, it may be possible that some of the misclassified instances are used for retraining, which is the reason for inferior performance. Above two experiments use Strong+Weak Teacher active learning system.

Go to Table of Contents