Supplementary Materials for MultiSports: A Multi-Person Video Dataset of Spatio-Temporally Localized Sports Actions

Yixuan Li Lei Chen Runyu He Zhenzhi Wang Gangshan Wu Limin Wang*
State Key Laboratory for Novel Software Technology, Nanjing University, China

1. More Dataset Details

1.1. Train split vs. Validation split

In order to guarantee enough instances for each class despite the severely unbalanced distribution, we artificially split the instances into the training set and the validation set in Table 1. To avoid data leakage from the training set to the validation/testing set, we ensure that data from the same match should be used for only one purpose. In other words, clips in the validation set cannot come from the matches covered in the training set. Unless otherwise mentioned, we report the results trained on the training set and evaluated on the validation set.

1.2. Comparison with other type of Dataset

MEVA [4] is a new security dataset, whose data is from RGB and thermal IR cameras, UAV footage and GPS locations for the actors. It defines 37 activities (66 for *MultiSports*) with 17055 instances (37701 for *MultiSports*), where 29 activities are about person and 8 activities are about vehicle. The categories in this dataset are atomic, such as $person\_close\_trunk$ and $person\_stand\_up$, which are different from our fine-grained and complicated sports categories. What's more, most of the categories in MEVA are daily actions, whose deformation and displacement are not large. Although it is a multi-person dataset, we believe our *MultiSports* can bring new challenges different from MEVA.

2. More Ablation Study

How the well-defined and high quality temporal boundary help? We add some temporal noise to the train set GT. For a L-frame length instance, we randomly choose a new length new_L from (1, L) and then the start point offset from (0, L-new_L). We sample the new annotation from the original. Other settings are kept the same. From the Table 2, we find the performance is much worse without well-defined temporal boundaries. It can conclude that our *MultiSports* has well-defined and high quality temporal annotations, which can help improve the performance and promote the algorithms to localize the boundary more accurately.

3. Method Details

ROAD [17] is a deep-learning framework for real-time action localisation and classification. It adopts SSD [14] to regress and classify action detection boxes in each frame independently, which does not utilize temporal information. Then, the frame detections are linked into action tubes by an online algorithm. Here we use the python linking code provided by MOC [11] instead of the original MATLAB code. Following the settings of ROAD on UCF101-24 [18], we use an ImageNet pre-trained VGG16 [16] network. We first try an initial learning rate of 1e-4 as their setting on UCF101-24, but the loss diverges into infinity after 20 iterations. The reported experiment on our *MultiSports* adopts an initial learning rate of 1e-5. We use SGD optimizer and the learning rate is reduced to its $\frac{1}{10}$ after 30000, 60000 iterations, which is the same as their practice on UCF101-24. The maximum iteration number is 120000.

YOWO [10] is a frame-level action detector with two branches. A 2D-CNN branch extracts the spatial features of the key frame while a 3D-CNN branch extracts spatio-temporal features of the key frame and the previous n (n=16) frames. Then, the features of two branches are fused by a channel fusion and attention mechanism(CFAM) module and finally passed to a convolution layer to predict the action class and bounding box in Yolov2 [15] manner. Finally, the frame detections are linked into action tubes by a dynamic programming algorithm. Note that the linking algorithm in YOWO is trimmed, thus we use the same linking algorithm as MOC on *MultiSports*. We use 2D Darknet-19 backbone pretrained on PASCAL VOC [7] and 3D ResNeXt-101 backbone pre-trained on Kinetics [1]. To utilize multiple GPUs, we modified the batch size to 80 and the initial learning rate to 8e-4. Following the training strategy of YOWO on UCF101-24 [18], we adopt SGD opti-

*Corresponding Author (lmwang@nju.edu.cn).

	Volleyball	Football	Basketball	Aerobic	All
instance ratio	3549:1294	6144:2153	4532:1715	4197:1415	18422:6577
clip ratio	402:130	402:132	379:147	391:146	1574:555
competition ratio	32:11	36:12	34:14	23:8	125:45

Table 1. Train split vs Validation split

Method	GT Noise	MultiSports		
		F@0.5	V@0.2	V@0.5
MOC (K=7)	✔	13.71	8.59	0.63
MOC (K=7)	✗	22.51	12.13	0.77
SlowOnly Det., 4×16	✔	12.60	8.98	3.05
SlowOnly Det., 4×16	✗	16.70	15.71	5.50

Table 2. Exploration on the effect of the temporal boundary noise.

mizer and the learning rate is reduced to its $\frac{1}{2}$ after 30000, 40000, 50000, 60000 iterations. The epoch maximum is set to 5. Note that YOWO only estimates performance on the frames having annotations, thus frame-mAP we report on UCF101-24 is much lower than in the original paper.

MOC [11] is an anchor-free tubelet-level action detector with three branches, which firstly takes K frames as input, then outputs K frame tubelet results and finally links these tubelets into tubes with a common matching strategy. We use DLA34 [21] as the backbone network, which is pre-trained on COCO [13]. Following the training strategy of MOC on UCF101-24 [18], we use the Adam optimizer with the learning rate 5e-4. The learning rate is reduced to its $\frac{1}{10}$ after epoch 6 and 8. The epoch maximum is set to 12.

SlowFast Det. [8] firstly uses a person detector on the key frame to localize for region proposal. Then, each 2D RoI at the key frame is extended into a 3D RoI by replicating it along the temporal dimension. Finally, it extracts RoI features from the backbone features for predicting category. The person detector is a Faster R-CNN with a ResNeXt-101-FPN [20, 12] backbone, which is pre-trained on ImageNet [6] and the COCO human keypoint images [13]. The backbone is the variant of SlowFast or SlowOnly, which sets the spatial stride of res_5 to 1 and uses a dilation of 2 for its filters. Note that we use the code in MMAction2 [3]. The results on AVA [9] and our *MultiSports* in the paper are all produced by it. We use the pre-computed proposals for AVA from previous work [8, 19]. Following previous work [8, 19], we fine-tune the person detector on our *MultiSports* with MMDetection [2]. We use the SGD optimizer with the learning rate 0.0025 and finetune 2 epochs on our *MultiSports*. The person detector produces 96.16 AR@100 on our *MultiSports* validation set. The detected boxes with confidence of > 0.9 are selected for action detection on both datasets. Our backbones are based on ResNet50, which are pre-trained on Kinetics-400 [1]. The $T \times \tau$ is set to 4×16. The α is set to 8 for SlowFast. We use a step-wise learning rate, reducing the learning rate $10\times$ after epoch 6 and 7. We train for 8 epochs with a linear warm-up for the first 5 epochs, where the result is similar with that of training 20 epochs and a lot of training time is saved. The initial learning rate is set to 0.1125 for SlowFast and 0.2 for SlowOnly. SlowFast and Slowonly Det. use the same link algorithm as MOC.

4. Error Analysis

4.1. Error Tree

To further understand the difficulty in our *MultiSports* dataset, we classify the detection errors into 10 different categories in a tree structure as shown in Figure 1 (code in $VideomAP\_error.py$), which are:

- E_R (Errors of repeated detections): a detection result that has tubelet IoU larger than a threshold and the right action class with some ground-truth tubelets, but the ground-truths have been matched by other detection results before with a confidence score larger than it.
- E_N (Errors of not matched): a detection result that has no intersection with any ground-truth tubelets of any class, indicating there should be no detection results but it appears out of thin air.
- E_L (Errors of spatial localization): a detection result that has the same action class and temporal IoU larger than a threshold with some ground-truth, but it has a low average spatial bounding box IoU in the area of the temporal intersection of ground-truth tubelets and it so that a lower tubelet IoU than the required threshold.
- E_C (Errors of classification): a detection result that has the tubelet IoU larger than a threshold with a ground-truth, but its action class is not the same with the ground-truth's class.
- E_T (Errors of temporal localization): a detection result that has the same action class and average spatial bounding box IoU larger than a threshold with some ground-truth in the area of the temporal intersection of ground-truth tubelets and it, but low temporal IoU

For each detected tubelet d_i from a sorted list by descending order of confidence score of class c.
Notation: **th:** threshold**; th_t :** the square root of **th**; **th_s:** the square root of **th; GT(c):** set of ground-truths of class c; **dupGT(c):** copy of **GT(c)** ; **GT(others) :** set of all ground-truths that not in class c; **GT(all):** set of all ground-truths; **T_IoU :** the temporal domain IoU; **S_IoU:** the average of the IoU between the overlapping frames; **tubelet_IoU: T_IoU* S_IoU.**

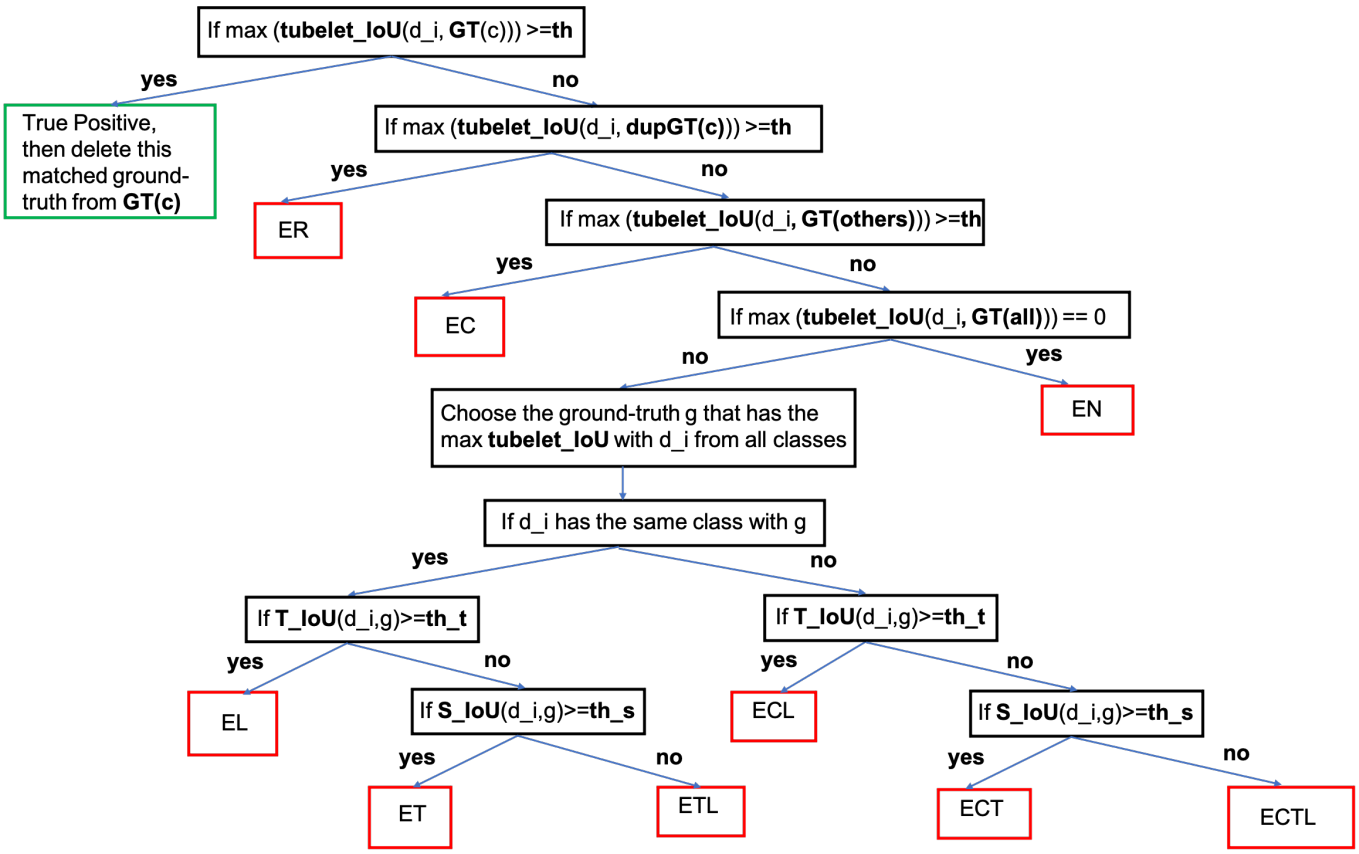


Figure 1. Error Tree

with ground-truths so that a lower tubelet IoU than the required threshold.

- $E_{C\&T}$ (Errors of both classification and temporal localization): a detection result that has average spatial bounding box IoU larger than a threshold with some ground-truth tubelets in the area of the temporal intersection of ground-truth tubelets and it, but both low temporal IoU and wrong action class.

- $E_{C\&L}$ (Errors of both classification and spatial localization): a detection result that has temporal IoU larger than a threshold with some ground-truth tubelets, but both wrong action class and low average spatial bounding box IoU with some ground-truth in the area of the temporal intersection of ground-truth tubelets and it.

- $E_{T\&L}$ (Errors of both temporal and spatial localization): a detection result in which we first select the ground-truth tubelet from all action classes that has the maximum tubelet IoU with the detection result, then we find they share the same action class, but both temporal IoU and average IoU of spatial bounding boxes lower than a threshold.

- $E_{C\&T\&L}$ (Errors of classification, temporal and spatial localization): a detection result that has some intersection with some ground-truth tubelets, which is different with EN, but wrong action class and both the temporal and average bounding box IoU lower than a threshold.

- E_M (Errors of missed detections): ground truth tubelets that have not been matched by any detection results.

4.2. More Visualization of Error Analysis

As shown in Figure 2, we collect more visualizations of MOC(K=11) as a supplementary of Figure 7 in our paper.

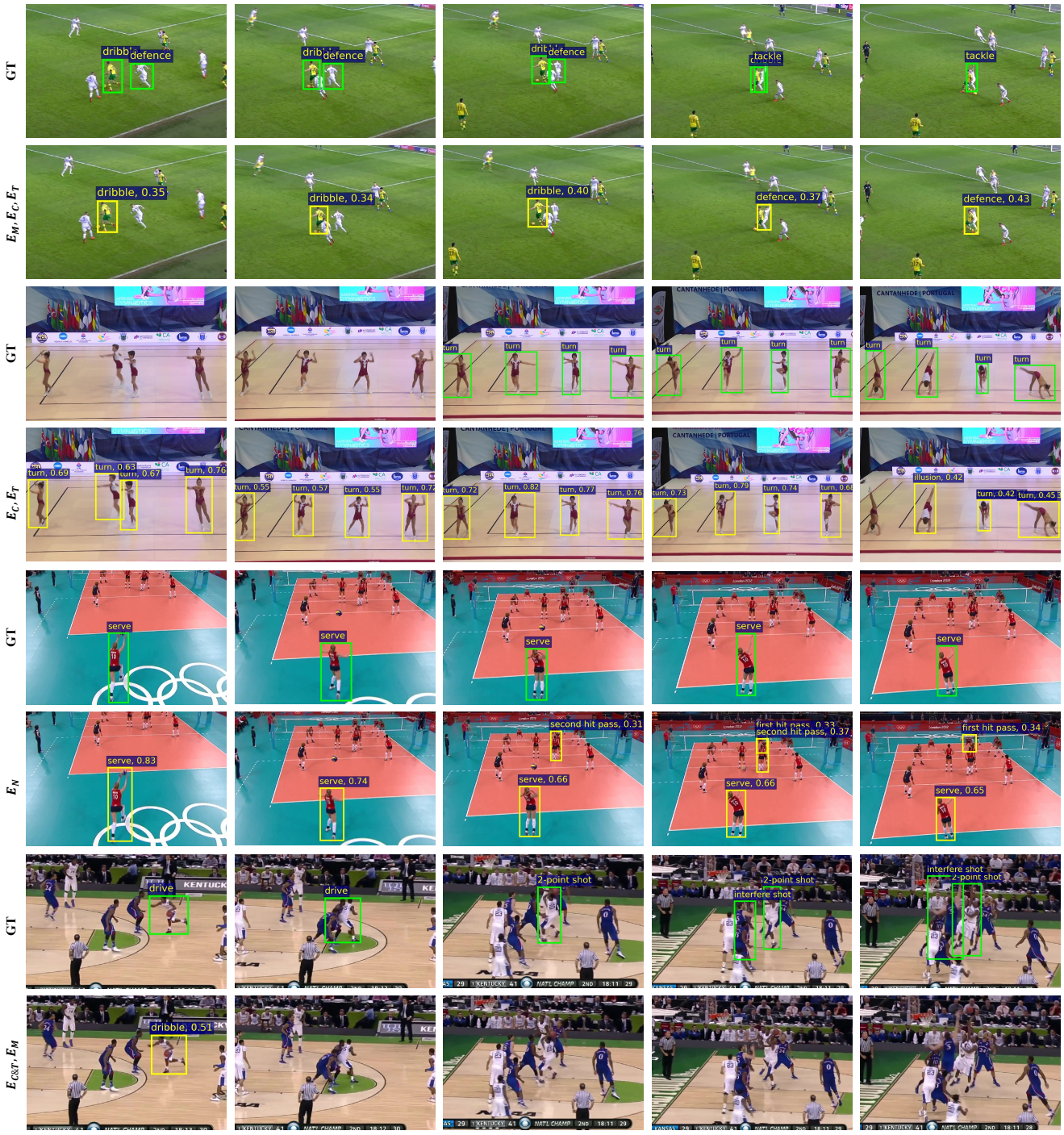
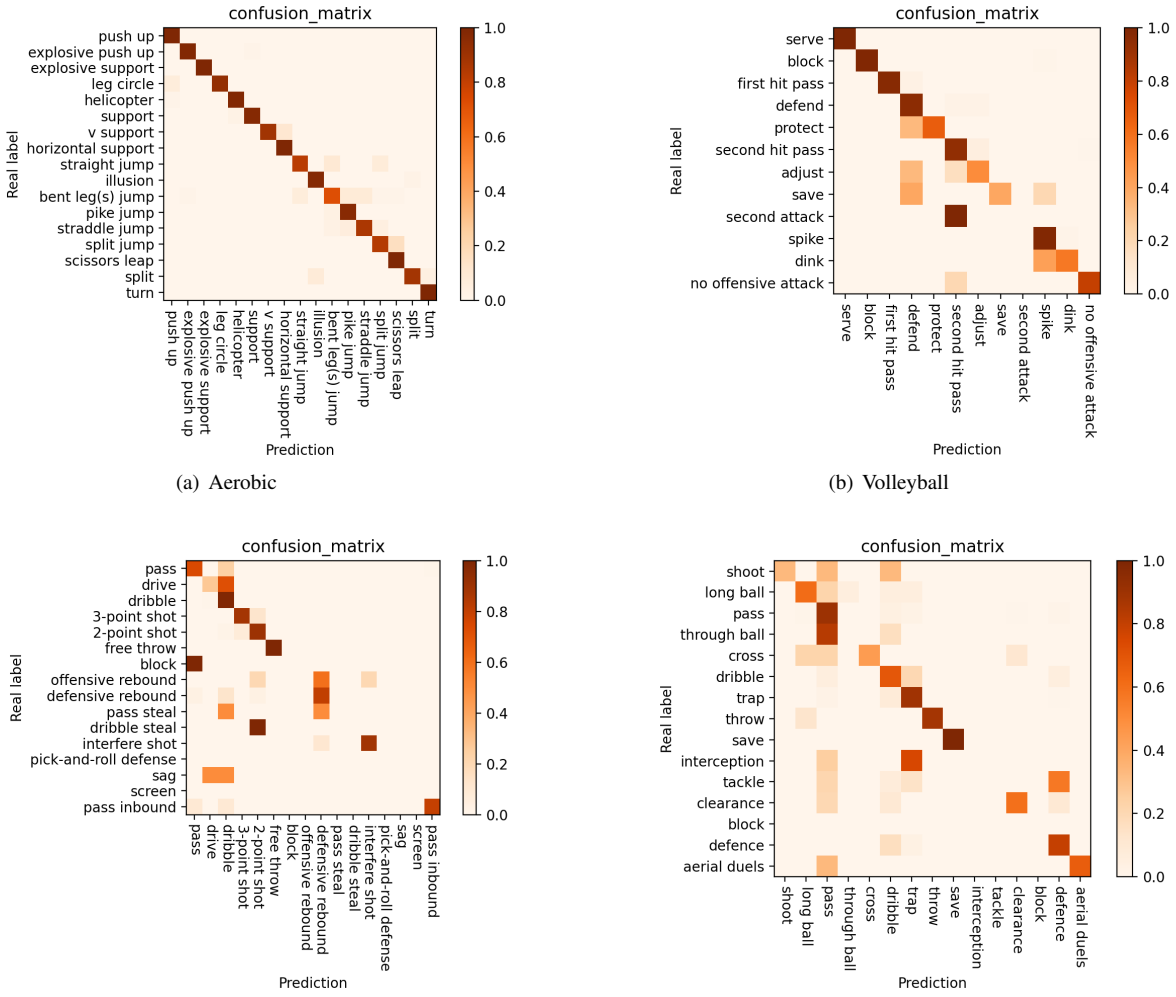


Figure 2. More detailed visualizations on our *MultiSports* dataset with our novel error categories of video-mAP. Green boxes are the ground-truths. Yellow boxes are the detections. 1st and 2nd row: E_M: missed detection of defence; E_C: tackle is misclassified as defence; E_T: dribble has inaccurate action ending boundary. 3rd and 4th row: E_C: turn is misclassified as illusion in the last picture in 4th row; E_T: turn has inaccurate action boundary. 5rd and 6th row: E_N: detection results contain that athletes actually doing none of sports actions but the model identifies first hit pass and second hit pass for them. 7rd and 8th row: $E_{C\&T}$: drive is misclassified as dribble and also has inaccurate action boundary; E_M: missed detections of interfere shot and 2-point shot.

4.3. Confusion Matrix

We draw the confusion matrices of the predictions which are classified into AP and E_C in Figure 3. We observe that



(a) Aerobic (b) Volleyball

(c) Basketball (d) Football

Figure 3. Confusion Matrix of SlowFast Det. on different sports.

the aerobic performs best because its categories relate only to individual actors. Actions having similar motions but different spatio-temporal contexts tend to confuse. For example, 1) drive vs. dribble in basketball, drive emphasizes on breaking through defender and being closer to the basket, which needs to model person-person interaction and spatial localization; 2) through ball vs. pass in football, through ball will break through the opponent's line of defense and be passed in front of the teammate, which needs long-term temporal modeling and reasoning. 3) offensive rebound vs. defensive rebound, the difference is whether the offensive player or defensive player gains control of the ball; 4) defend vs. protect in volleyball, we need to focus on whether the ball was blocked back or was spiked by an opponent several frames earlier.

5. Annotation Documentation

5.1. Aerobic Gymnastics

There are four groups of difficulty elements in aerobic gymnastics, namely dynamic strength, static strength, jumps & leaps, and balance & flexibility. We pick out 21 elements to form the aerobic categories of our *MultiSports*. The following is a detailed definition of these categories, a simplified version of the definition in [5].

Group A: Dynamic Strength. All elements in Group A ending in a split position, must have both hands on each side of the body on the floor.

- **Push up**: Starting and/or finishing: one or both hands are in contact with the floor, elbows extended. Shoulders must be parallel to the floor; head in line with the spine and pelvis tucked with abdominal muscles contracted. Flexion of elbows: All push-ups must have, at the end of the downwards phase, a maximum distance

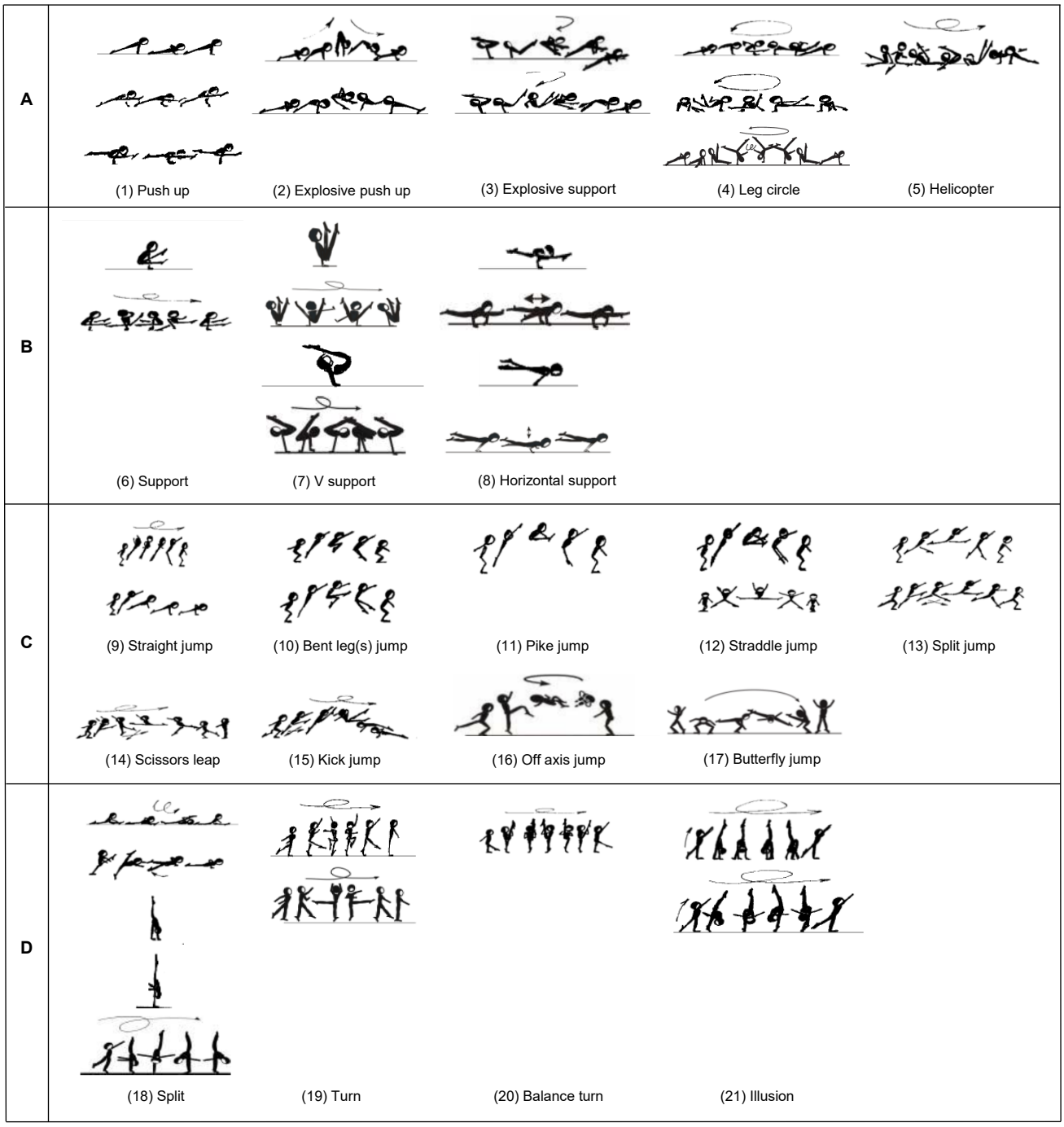


Figure 4. Diagrams of each difficulty element in aerobic gymnastics.

of 10cm from the chest to the floor. The downward and/or the upward phase of a push up must be controlled with shoulders parallel to the floor. Lateral and Hinge push up, 4 phases have to be shown. Wenson push up: one leg on the upper part of the arm (Triceps) of the same side.

- **Explosive push up**: 1) A Frame: Pike position in the airborne phase (60° between trunk and legs). 2) Cut: While airborne, the legs straddle sideways and forward to land extended in rear support, feet lifted off the floor during the skill.

- **Explosive support**: Back support on the floor, back parallel to the floor, extending the legs upward and for-

ward with a flight phase. Impulse from High V support position, airborne phase and landing to push up or split position.

- **Leg circle**: The starting position must be from free front support on both hands; the hips must be lifted and extended during the full rotation. Feet may not touch the floor before the completion of the circle. 1) Leg circle: the hips must be lifted and extended. 2) Flair: legs straddle, the hips must be lifted and extended during the full rotation. Feet may not touch the floor before the completion of the circle.

- **Helicopter**: After alternative leg circles, legs close to the chest, body alignment on the upper back (feet off the floor). The legs are extended upward and forward. ½ twist initiated from the feet is made to land in push up or wenson or split.

Group B: Static Strength. These elements demonstrate isometric strength and must be held for 2 seconds. In the case of turns in support, the support must be held for 2 seconds either at the start, during or end of the turn. The body is fully supported by one or both arms and only the hands are in contact with the floor. Feet and/or hips must not touch the floor during the whole skill. While in support, the hands must be flat on the floor.

- **Support**: 1) Straddle support: Legs must be straight parallel to the floor in Straddle position (90°minimum). 2) L support: Legs must be straight together and parallel to the floor.

- **V support**: 1) Straddle V support: Hips are flexed and legs straddled 90° open and vertical, minimum width 90°. 2) V support: Hips are flexed and legs are together vertical. 3) High V support: The back is parallel to the floor.

- **Horizontal support**: 1) Wenson support: the body is extended parallel to the floor, one leg supported on the upper part of the Triceps. 2) Planche: the body is supported on both hands with straight arms, not more than 20° above parallel.

Group C: Jumps & Leaps. All jumps and leaps must demonstrate explosive power and maximum amplitude. All jumps that can be performed from 1 foot or two feet will be considered as the same element and will receive the same value. This applies also for landing. Take off preparation: head, shoulder, chest, hips, knees, feet must in the same direction. Body shape while airborne must be clearly recognizable. Body and legs must be tight and straight, with head in line with the spine. **Landing Positions**: 1) Standing: Landing on one foot or two feet must be in a vertical position, with bend leg(s) before finishing in perfect alignment. 2) In push up: both hands and supporting feet must land at the same time in a controlled manner. 3) In split: must land from airborne phase to split form with both hands on each side of the body on the floor. 4) In frontal split: must land from airborne phase to frontal split form, both hands in front of the body.

- **Straight jump**: The body is in extended alignment, the pelvis is fixed – 2 different kinds of jumps and leaps: 1) Vertical: All air turns, Free fall. 2) Vertical to Horizontal: Gainer.

- **Bent leg(s) jump**: 1) Tuck: Both legs are lifted close to the chest with knees bent. 2) Cossack: After takeoff, the body shows a pike shape, legs together parallel to the floor or higher, one leg straight, one leg bent. The angle between the trunk and legs: not be more than 60°. The angle at the knee joint may not be more than 60°.

- **Pike jump**: After takeoff, the body shows a pike shape, legs together and straight, parallel to the floor or higher. The angle between the trunk and legs may not be more than 60°.

- **Straddle jump**: 1) Straddle: Legs are lifted in straddle position (minimum 90° angle), parallel to the floor or higher, arms and trunk extended over them. The angle between the trunk and legs may not be more than 60°. 2) Frontal split: Legs are fully abducted laterally (right and left) frontal (180°) with the upright upper body.

- **Split jump**: 1) Split: Legs are fully stretched front and back in sagittal split (180°) with the upright upper body. 2) Switch: After takeoff, the leading leg must be parallel to the floor and switch with the rear leg to show a split (180°) in the air.

- **Scissors leap**: The leading leg must be parallel to the floor and switches forward with 1/2 turn (180°).

- **Kick jump**: The leading leg must be parallel to the floor and switches forward.

- **Off axis jump**: A one-foot take off, kicking the free leg (bend or straight) upward diagonally. While airborne, the body inclines backward to be out of axis with a longitudinal rotation(s) in tuck or straight position, arms close to the chest. Landing in 1 foot/feet together or in split.

- **Butterfly jump**: A one-foot take off, kicking the free leg backward to lift the body upward. While airborne, legs fly open in straddle (or feet together) with the body in a horizontal position (with or without longitudinal rotation(s). Landing on one leg.

Group D: Balance & Flexibility.

- **Split**: Legs must be straight, in line, showing 180°. In Vertical Split: supporting leg must be in vertical position.
- **Turn**: All exercises requiring turns must demonstrate complete rotations on the ball of the foot. Turns are completed when the heel of the turning foot touches the floor.
- **Balance turn**: A Balance turn where one leg is lifted to either in sagital or frontal balance and is supported by one hand.
- **Illusion**: Starting position of illusion: head, shoulder, chest, hips, knees, toes must be in alignment. A full split (180°) must be shown during the movement.

For the temporal definition, strictly follow the diagrams in Figure 4 (quoted from [5]) to determine the starting and ending of actions, except for the following situations: when an athlete's action is not in place or is completely blocked by other athletes.

5.2. Volleyball

- **Serve**: Send the ball over the net from behind the end line to start a new round. **Start**: The ball leaves the player's hand. **End**: If the player takes off, any foot touches the ground. Otherwise, the upper arm of the serving arm is below the horizontal plane.
- **Block**: Deflect the ball coming from an attacker on the net. The one that doesn't take off is not considered a block. The one that takes off but doesn't touch the ball is considered a block. **Start**: Any foot leaves the ground. **End**: Any foot touches the ground.
- **First Hit Pass**: Receive the serve. The player can receive the ball overhand, one-hand or underhand. **Start**: If overhand, the player raises any hand over the chest. If one-hand, the player's arm begins to stretch out. If underhand, the player begins to hold hands together. **End**: If overhand, the player puts any hand below the chest. If one-hand, the player's hitting-ball arm relaxes. If underhand, the player's hands loose.
- **Defend**: Receive the ball from the opposite side except for the serve. The player can receive the ball overhand, one-hand or underhand. **Start**: If overhand, the player raises any hand over the chest. If one-hand, the player's arm begins to stretch out. If underhand, the player begins to hold hands together. **End**: If overhand, the player puts any hand below the chest. If one-hand, the player's hitting-ball arm relaxes. If underhand, the player's hands loose.
- **Protect**: Receive the ball returned by the block. The player can receive the ball overhand, one-hand or underhand. **Start**: If overhand, the player raises any hand over the chest. If one-hand, the player's arm begins to stretch out. If underhand, the player begins to hold hands together. **End**: If overhand, the player puts any hand below the chest. If one-hand, the player's hitting-ball arm relaxes. If underhand, the player's hands loose.
- **Second Hit Pass**: The second overhand pass to organize the offense. **Start**: The player raises any hand over the chest. **End**: The player puts any hand below the chest.
- **Adjust**: For the second touch, due to the inadequacy of first hit pass or defending or protecting, the player has to adjust the ball underhand to organize offense. **Start**: The player begins to hold hands together. **End**: The player's hands loose.
- **Save**: Due to the poor first hit pass or defending or protecting, the route of the ball changes dramatically. The actor can't second hit pass overhand or adjust underhand to organize offense but uses one hand or both hands to reach the ball to prevent the ball from landing directly. **Start**: If one-hand, the player's arm begins to stretch out. Otherwise, the player begins to hold hands together. **End**: If one-hand, the player's hitting-ball arm relaxes. Otherwise, the player's hands loose.
- **Second Attack**: For the second touch, a direct attack by the setter. **Start**: Any foot leaves the ground. **End**: Any foot touches the ground.
- **Spike**: Slam the ball over the net into the opposing court. **Start**: Any foot leaves the ground. **End**: Any foot touches the ground.
- **Dink**: Lightly tap the ball over the net to an area on the opponent's side of the court that is not being guarded or occupied by a defensive player. **Start**: Any foot leaves the ground. **End**: Any foot touches the ground.
- **No Offensive Attack**: For the second or third touch, the ball is passed over the net non-aggressively, because of the bad first hit pass or defending or protecting. The actor can push the ball overhand, pass the ball underhand or tap the ball from a position below the net with one hand, where the actor doesn't take off. **Start**: If overhand, the player raises any hand over the chest. If underhand, the player begins to hold hands together. Otherwise, the upper arm of hitting the ball arm is above the horizontal plane. **End**: If overhand, the player puts any hand below the chest. If underhand, the player's hands loose. Otherwise, the upper arm of hitting the ball arm is below the horizontal plane.

5.3. Football

- **Shoot**: Hit the ball in an attempt to score a goal. Feet, torso and head are all allowed. **Start**: If using torso and head: if the player takes off, any part of the body leaves the ground (such as a foot); if the player does not take off, the player stands firmly and prepares to touch the ball. If using feet, the ball-controlling foot leaves the ground. **End**: If using torso and head: if the player takes off, any part of the body touches the ground (such as a foot); if the player does not take off, stand firmly after touching the ball. If using feet, the ball-controlling foot touches the ground.

- **Long Ball**: Middle and long distance (over 30 meters) pass. Feet, torso and head are all allowed. **Start**: If using torso and head: if the player takes off, any part of the body leaves the ground (such as a foot); if the player does not take off, the player stands firmly and prepares to touch the ball. If using feet, the ball-controlling foot leaves the ground. **End**: If using torso and head: if the player takes off, any part of the body touches the ground (such as a foot); if the player does not take off, stand firmly after touching the ball. If using feet, the ball-controlling foot touches the ground.

- **Pass**: Short distance (within 30 meters) pass. Feet, torso and head are all allowed. **Start**: If using torso and head: if the player takes off, any part of the body leaves the ground (such as a foot); if the player does not take off, the player stands firmly and prepares to touch the ball. If using feet, the ball-controlling foot leaves the ground. **End**: If using torso and head: if the player takes off, any part of the body touches the ground (such as a foot); if the player does not take off, stand firmly after touching the ball. If using feet, the ball-controlling foot touches the ground.

- **Through Ball**: A pass that can clearly break through the opponent's line of defense and has a penetrating effect. At least one defensive player is passed. The ball is passed in front of the player's teammate. In other words, the player should pass the ball to where his running teammate is going to be. Feet, torso and head are all allowed. **Start**: If using torso and head: if the player takes off, any part of the body leaves the ground (such as a foot); if the player does not take off, the player stands firmly and prepares to touch the ball. If using feet, the ball-controlling foot leaves the ground. **End**: If using torso and head: if the player takes off, any part of the body touches the ground (such as a foot); if the player does not take off, stand firmly after touching the ball. If using feet, the ball-controlling foot touches the ground.

- **Cross**: A medium-to-long-range pass from a wide area of the field towards the centre of the field near the opponent's goal. Provide direct or indirect shooting opportunities for offensive players. Feet, torso and head are all allowed. **Start**: If using torso and head: if the player takes off, any part of the body leaves the ground (such as a foot); if the player does not take off, the player stands firmly and prepares to touch the ball. If using feet, the ball-controlling foot leaves the ground. **End**: If using torso and head: if the player takes off, any part of the body touches the ground (such as a foot); if the player does not take off, stand firmly after touching the ball. If using feet, the ball-controlling foot touches the ground.

- **Dribble**: Have control over the ball for a period of time and distance. **Start**: At the first touch with the ball, the ball-controlling foot leaves the ground. **End**: At the last touch with the ball, the ball-controlling foot touches the ground.

- **Trap**: Use effective parts of the body to adjust the ball, including the speed and position of the ball. Feet, torso and head are all allowed. **Start**: If using torso and head: if the player takes off, any part of the body leaves the ground (such as a foot); if the player does not take off, the player stands firmly and prepares to touch the ball. If using feet, the ball-controlling foot leaves the ground. **End**: If using torso and head: if the player takes off, any part of the body touches the ground (such as a foot); if the player does not take off, stand firmly after touching the ball. If using feet, the ball-controlling foot touches the ground.

- **Throw**: The player throws the ball from out of the field and the goalkeeper throws the ball. **Start**: Upper arms swing forward. **End**: Upper arms are below the horizontal plane.

- **Save**: The goalkeeper uses his body parts (except his feet) to destroy the ball that is threatening to the goal. **Start**: After the ball is shot, the goalkeeper begins to move. **End**: Any part of the body touches the ground.

- **Interception**: The defensive player consciously destroys the ball on the opponent's pass route. Feet, torso and head are all allowed. **Start**: If using torso and head: if the player takes off, any part of the body leaves the ground (such as a foot); if the player does not take off, the player stands firmly and prepares to touch the ball. If using feet, the interception foot leaves the ground. **End**: If using torso and head: if the player takes off, any part of the body touches the ground (such as a foot); if the player does not take off, stand firmly after touching the ball. If using feet, the interception foot touches the ground.

- **Tackle**: The defensive player snatches the ball under the control of the offensive player. **Start**: the tackling foot leaves the ground. **End**: the tackling foot touches the ground.

- **Clearance**: The defensive player destroys the ball in the backfield in order to gain the initiative in time and space. The main difference between clearance and long ball/ball is that long ball/ball aims at some player but clearance is aimless. Feet, torso and head are all allowed. **Start**: If using torso and head: if the player takes off, any part of the body leaves the ground (such as a foot); if the player does not take off, the player stands firmly and prepares to touch the ball. If using feet, the clearance foot leaves the ground. **End**: If using torso and head: if the player takes off, any part of the body touches the ground (such as a foot); if the player does not take off, stand firmly after touching the ball. If using feet, the clearance foot touches the ground.

- **Block**: Intentionally destroy the opponent's threatening shot or block the opponent's shooting angle. The goalkeeper blocked the ball with his foot. Feet, torso and head are all allowed. **Start**: If using torso and head: if the player takes off, any part of the body leaves the ground (such as a foot); if the player does not take off, the player stands firmly and prepares to touch the ball. If using feet, the blocking foot leaves the ground. **End**: If using torso and head: if the player takes off, any part of the body touches the ground (such as a foot); if the player does not take off, stand firmly after touching the ball. If using feet, the blocking foot touches the ground.

- **Defence**: The defensive player approaches the player, of whom the ball is under the control, to make restriction and interference. **Start**: The defender is shorter than 1.2 meters from the offensive player who is controlling the ball. **End**: 1) the offensive player passes the ball out or the ball is gained by other defensive players. 2) this defender begins to tackle. 3) this defender is longer than 1.2 meters from the offensive player who is controlling the ball.

- **Aerial duels**: Two or more people compete for the high-altitude ball in order to obtain the ball, where all people are annotated. If the player does not take off, it is not considered aerial duels. Note that the player who has an obvious purpose, such as clearance and pass, is annotated that action. **Start**: Any part of the body leaves the ground (such as a foot). **End**: Any part of the body touches the ground (such as a foot).

5.4. Basketball

- **Pass**: The player moves the ball to the teammate. **Start**: The player begins to push the ball outwards with his arms. **End**: The ball leaves both hands of the player.

- **Drive**: The player, who controls the ball, gets rid of the defense by passing the defensive player or stopping suddenly during the movement. The aim is to get closer to the basket and create a space that is conducive to shooting. The next step is usually to shoot, layup, or pass the ball to teammates. **Start**: At the first touch with the ball, the hand presses the ball down. **End**: At the last touch with the ball, the ball is bounced into the hand.

- **Dribble**: The player slaps the ball bounced from the ground continuously while on the spot or on the move. **Start**: At the first touch with the ball, the hand presses the ball down. **End**: At the last touch with the ball, the ball is bounced into the hand.

- **3-point Shot**: Shoot from beyond the three-point line. **Start**: The player raises the shooting hand over the chest. **End**: If the player takes off, any part of the body touches the ground (such as a foot). Otherwise, the player puts the shooting hand below the chest.

- **2-point Shot**: Shoot from within the three-point line. **Start**: The player raises the shooting hand over the chest. **End**: If the player takes off, any part of the body touches the ground (such as a foot). Otherwise, the player puts the shooting hand below the chest.

- **Free Throw**: Unopposed attempts to score points by shooting from behind the free throw line. **Start**: The player raises the shooting hand over the chest. **End**: If the player takes off, any part of the body touches the ground (such as a foot). Otherwise, the player puts the shooting hand below the chest.

- **Block**: When the offense shoots, the defender successfully knocks the ball out as the ball goes up. **Start**: The defender raises the blocking hand over the chest. **End**: If the defender takes off, any part of the body touches the ground (such as a foot). Otherwise, the defender puts the blocking hand below the chest.

- **Offensive Rebound**: After a missed shot, the two sides compete for a rebound and the offensive player grabs it. **Start**: The player raises the grabbing-ball hand over the chest. **End**: If the player takes off, any part of the body touches the ground (such as a foot). Otherwise, the player catches the ball firmly.

- **Defensive Rebound**: After a missed shot, the two sides compete for a rebound and the defensive player grabs it. **Start**: The player raises the grabbing-ball hand over the chest. **End**: If the player takes off, any part of the body touches the ground (such as a foot). Otherwise, the player catches the ball firmly.
- **Pass Steal**: The defensive player intercept the ball in the process of passing, which is not under the control of offensive player. **Start**: The defender's stealing-ball hand begins to stretch out. **End**: 1) Route of the ball changes; 2) The defender catches the ball firmly.
- **Dribble Steal**: The defensive player steals the ball under the control of offensive player. **Start**: The defender's stealing-ball hand begins to stretch out. **End**: The ball is out of the control of the offensive player who had the control.
- **Interfere Shot**: The defender interferes with the shot but does not touch the ball. **Start**: The defender raises the interfering hand over the chest. **End**: If the defender takes off, any part of the body touches the ground (such as a foot). Otherwise, the defender puts the interfering hand below the chest.
- **Pick-and-roll Defense**: In pick-and-roll, the defender of the offensive ball-controlling player is blocked by the teammate of this offensive player. **Start**: The defender has physical contact with the offensive screening player. **End**: The defender does not have physical contact with the offensive screening player.
- **Sag**: The defender gives up the offensive player he is responsible for and turns to defend the offensive ball-controlling player. **Start**: The defender consciously approach the offensive ball-controlling player. **End**: 1) the offensive player passes or shoots the ball; 2) this defender is broken through; 3) this defender gives up.
- **Screen**: In pick-and-roll, the offensive player uses his body to set a pick for his ball-controlling teammate. **Start**: Both feet of the offensive player touches the ground. **End**: Any foot of the offensive player is ready to leave the ground completely. Small range movement is not considered the end.
- **Pass-inbound**: The player passes the ball from the boundary lines to restart the play. **Start**: The player begins to push the ball outwards with his arms. **End**: The ball leaves both hands of the player.
- **Save**: The player gets back the ball that is about to go out of bounds. **Start**: The player begins to push the ball outwards with his arms. **End**: The ball leaves both hands of the player.
- **Jump Ball**: A method used to begin or resume the play. Two opposing players attempt to gain control of the ball after an official tosses it into the air between them, where both players are annotated. **Start**: The player raises the grabbing-ball hand over the chest. **End**: Any part of the body touches the ground (such as a foot).

References

[1] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR*, pages 4724–4733, 2017. 1, 2

[2] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. Mmdetection: Open mmlab detection toolbox and benchmark. *CoRR*, abs/1906.07155, 2019. 2

[3] MMAction2 Contributors. Openmmlab's next generation video understanding toolbox and benchmark. https://github.com/open-mmlab/mmaction2, 2020. 2

[4] Kellie Corona, Katie Osterdahl, Roderic Collins, and Anthony Hoogs. MEVA: A large-scale multiview, multimodal video dataset for activity detection. In *WACV*, pages 1059–1067, 2021. 1

[5] Federation Internationale de Gymnastique. Aerobic gymnastics-code of points. *FIG Aerobic Gymnastics FIG Executive Committee*, 2017. 5, 8

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 2

[7] Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vis.*, pages 98–136, 2015. 1

[8] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, pages 6201–6210, 2019. 2

[9] Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. AVA: A video dataset of spatio-temporally localized atomic visual actions. In *CVPR*, pages 6047–6056, 2018. 2

[10] Okan Köpüklü, Xiangyu Wei, and Gerhard Rigoll. You only watch once: A unified CNN architecture for real-time spatiotemporal action localization. *CoRR*, abs/1911.06644, 2019. 1

[11] Yixuan Li, Zixu Wang, Limin Wang, and Gangshan Wu. Actions as moving points. In *ECCV*, pages 68–84, 2020. 1, 2

[12] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 936–944, 2017. 2

[13] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, pages 740–755, 2014. 2

[14] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: single shot multibox detector. In *ECCV*, pages 21–37, 2016. 1

[15] Joseph Redmon and Ali Farhadi. YOLO9000: better, faster, stronger. In *CVPR*, pages 6517–6525, 2017. 1

[16] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 1

[17] Gurkirt Singh, Suman Saha, Michael Sapienza, Philip H. S. Torr, and Fabio Cuzzolin. Online real-time multiple spatiotemporal action localisation and prediction. In *ICCV*, pages 3657–3666, 2017. 1

[18] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012. 1, 2

[19] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krähenbühl, and Ross B. Girshick. Long-term feature banks for detailed video understanding. In *CVPR*, pages 284–293, 2019. 2

[20] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, pages 5987–5995, 2017. 2

[21] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *CVPR*, pages 2403–2412, 2018. 2