Mask R-CNN and Optical Flow Based Method for Detection and Marking of Handball Actions

Pobar, Miran; Ivašić-Kos, Marina

Mask R-CNN and Optical flow based method for detection and marking of handball actions

Miran Pobar
Department of Informatics
University of Rijeka
Rijeka, Croatia
mpobar@uniri.hr

Marina Ivašić-Kos
Department of Informatics
University of Rijeka
Rijeka, Croatia
marinai@uniri.hr

***Abstract*— To build a successful supervised learning model for action recognition a large amount of training data needs to be labeled first. Labeling is normally done manually and it is a tedious and time-consuming task, especially in the case of video footage, when each individual athlete performing a given action should be labeled. To minimize the manual labor, we propose a Mask R-CNN and Optical flow based method to determine the active players who perform a given action among all players presented on the scene. The Mask R-CNN is a deep learning object recognition method used for player detection and optical flow measures player activity. Combining both methods ensures tracking and labeling of active players in handball video sequences. The method was successfully tested on a dataset of handball practice videos recorded in the wild.**

Keywords— object detectors; sports scenes; Mask R-CNN, optical flow; action recognition database

I. INTRODUCTION

For a supervised machine learning task, such as classification, object detection and action recognition a prerequisite is a training dataset with labeled objects or action clips. Usually, it is easier to detect the discriminatory features or discover rules contained in the data when a large amount of data is used, so the learned model can better fit the data and better generalize when a larger number of examples are used. For example, the MS COCO dataset designed for object detection and segmentation that is frequently used in deep learning classification and object detection tasks contains over two million labeled objects in about 200 000 images [1].

There is no doubt that one of the key factors contributing to the great success of deep learning solutions for classification and object detection is the availability of very large learning datasets designed for these tasks. But, preparation of a learning dataset for some specific task is not an easy job. For example, during the annotation of the COCO image database over 22 working hours per 1,000 segmentations of object instances were needed [1].

In the case where video materials are used instead of images, preparation of learning dataset is even more demanding, since video combines the audio and temporal components with spatial image components. The annotator should view each video frame, determine the beginning and end of an action, cut clips and mark objects on frames. If the video contains multiple objects, each object should be marked in all frames. Also, if the video contains multiple persons and the goal is to recognize actions performed by people in videos, each person and all sequences of performed actions should be marked as well. Consequently, development of learning datasets for video analyses such as action detection and recognition is very slow, expensive, involves many working hours and is a long-lasting process.

In this paper, we propose a method for extracting and tagging actions in video clips in order to build a dataset for detection and recognition of actions in handball. Detection involves temporal and spatial segmentation of the person performing the action (i.e. bounding boxing), while recognition implies labeling the performed action (e.g. jumping, running, dribbling, etc.). The videos are captured in the wild during handball training sessions with many players on the field performing various actions. The recording was done under uncontrolled conditions, with cluttered background, indoor with complex lighting conditions and shadows, with multiple actors who dynamically change their positions and cover each other. Therefore, standard and simple methods for person segmentation and mask detection such as background subtraction or Chroma keying are no longer a satisfactory solution.

Background subtraction and Chroma keying were successfully employed on early action recognition datasets such as KTH [2] and Weizmann [3] for segmentation of a single actor from the homogeneous background [4]. Scenes were recorded with a static camera in controlled conditions with one person performing an action over a homogeneous background, so the temporal segmentation could be done manually in a reasonable amount of time.

Today's datasets are far more complex, but labeling the video clips and marking objects with bounding boxes are mostly manual and even more dependent on human work. In order to employ many people as human annotators, most often crowdsourcing services such as Amazon Mechanical Turk are used. However, the problem of video annotating remains challenging even with crowdsourcing services, as it has been shown [5] that most crowdsourcing workers produce unreliable annotations that need to be thoroughly checked, and the cost and time spend are still significant especially if one takes into account that the databases are now very large.

To facilitate image and video labeling in a distributed environment, specialized online annotation tools were developed, e.g. LabelMe video [6]. Using this tool, the annotators can mark relevant actor shapes in certain key frames,

and then missing annotations between key frames are inserted using interpolation in conjunction with global motion estimation. The interpolated shapes across different frames can later be manually edited and corrected. An iterative approach is used in [7], where a classifier is used to assist in data annotation. The classifier is trained on a small number of labeled frames and is then used to classify/annotate the remaining data, which is then in turn used to improve the classifiers.

As an alternative to labeling large amounts of real-world data, a game engine is used in [8] to generate a synthetic dataset of action recognition videos. The synthetic data set was used to increase the actual set of data. The advantage of this approach is that large amounts of relatively diverse data can be generated procedurally without the need for manual segmentation of actors.

In this paper, we propose a pipeline to accelerate and facilitate building of a learning dataset for action recognition. The idea is to use the existing detection and tracking methods to determine active players in video sequences where specific activities are taking place, i.e. those who perform the relevant action. For these reasons, we propose the MOF method that combines the Mask R-CNN [9] for person detection and the optical flow method [10, 11] to detect changes in the position of the moving object within the bounding box, its speed and direction and deformation between frames.

The rest of the paper is organized as follows: Section II. provides the problem definition. Section III. describes the proposed MOF method for mapping actions in videos involving the use of Mask R-CNN object detector and optical flow method. In Section IV. the performance on a custom dataset consisting of indoor and outdoor handball scenes recorded during handball school are explained and discussed. The paper ends with a conclusion and the proposal for future research.

II. Problem Definition

Handball is an Olympic sport involving two 7-player teams playing in the hall on the handball court. The goal of the game is to hit the ball into the goal and score more goals than the opposing team to win. All players are free to move all over the field except in the space 6 m in front of both goals where the goalkeeper stands. Players quickly change their positions and roles in the game between offense or defense, combining different actions. The rules of the handball game are well-defined and prescribe permitted actions and techniques, but during training and school lessons coach usually modifies these rules to maintain a high activity level with fast technique change and more repetitions.

The video material we use is recorded during the school training sessions when about 15 players practice and repeat certain handball techniques.

The training consists of several sets of exercises that practice certain techniques or actions. For example, when practicing the shooting techniques, one player shoots the ball to the goal, the goalkeeper defends the goal, and the other students wait in the queue to perform the action or collect their balls around the court and return to the queue, Fig. 1. In that case, active players are the player who throws the ball to the goal and the other who is running to take the position, and the other players waiting in the queue are inactive.

Fig. 1. A typical training situation. Two players on the right are performing the current task, while the rest are queuing.

III. Proposed MOF method for marking handball actions

The aim of the proposed method is to determine the active players among others present in video sequences, who perform specific handball actions or technique.

The pipeline of the proposed method is shown in Fig. 2.

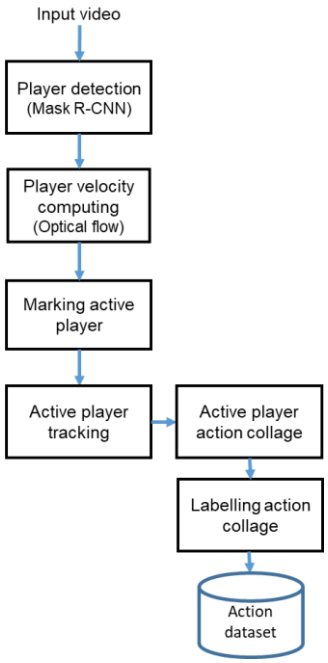

Fig. 2. A pipeline of MOF method for building handball action dataset.

The input is video sequence from video materials captured during the handball training. The video sequence is manually selected so it contains all the phases of a given action. Then, in each video sequence, all the players are detected and marked with a bounding box on each frame using Mask R-CNN. For each bounding box, optical flow is calculated considering the velocity component of the motion. Players who are moving faster in a given number of successive frames will have higher velocity values and will be marked as active in that video sequence. All occurrence of the player when is selected as active are stored in the collage that represents the action. The collage is then manually denoted by the name of the handball action it represents and added to the dataset.

A. *Player detection*

A convolutional neural network Mask R-CNN [9] is used to detect and localize players in video frames. The Mask R-CNN is a successor of Faster R-CNN [12] that follows a two-stage design. In first stage candidate object bounding boxes or regions of interest (RoI) are proposed and then a deep fully convolutional network is applied in a sliding window fashion to examine the probability whether it is an object class or a background. Then network generates bounding boxes and masks for all possible classes with the corresponding confidence values. In a parallel fully connected branch network predicts segmentation masks on each RoI to make the selection of boxes and object masks.

We have used the standard Resnet-101-FPN network configuration of Mask R-CNN with pre-trained parameters on the COCO dataset. In this experiment, we only consider bounding boxes that refer to the "person" class.

To reduce false positive person detection, we used detection confidence value and only selected those detections that had confidence above the set threshold. The threshold is experimentally set to 0.55 to ensure a good balance of high detection rates and low false positives. The R-CNN detection results for the "person" class are shown in Fig. 3.

Fig. 3. Results of Mask R-CNN person detection

B. *Player velocity computing and active player marking*

The detector captures the location of all the players present in the scene in bounding boxes but has no information about the movements or activities of the players.

It is expected that various actions in sports videos will be characterized by strong and sudden changes in velocity and appearance over time. Here we use the estimate of optical flow from time-varying image intensity to represent the information about speed and direction of movement of players. Movements of each point on the image plane produce a 2D path $x(t) \equiv (x(t), y(t))^T$ in camera-centered coordinates. The current direction is the velocity $dx(t)/dt$. For all visible surface points, the 2D velocities are often referred to as the 2D motion field.

The optical flow method by itself operates on sections of video frames without knowledge about object identities. Therefore, to measure the player activity we combine the information about player locations obtained with Mask RCNN with a measure of their movement speed represented by optical flow within the player bounding box.

The optical flow between consecutive video frames is estimated using the Lucas-Kanade method [11]. The Lucas-Kanade method first divides the original images into smaller sections and assumes a constant velocity in each section. The result is a vector field V of velocities of each image section, where at each point (x,y) the vector magnitude represents the movement speed and the angle represents the direction of movement.

A visualization of the optical flow vectors calculated on two video frames from the dataset is shown in Fig. 4. The direction and magnitude of optical flow at each point is represented by the direction and length of each arrow.

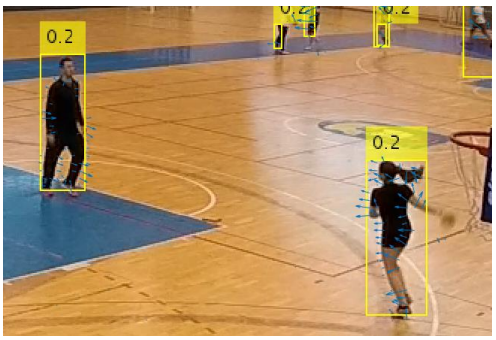

Fig. 4. Player bounding box (yellow) and optical flow vectors (blue).

Since video materials are recorded in real conditions, multiple players will move on the scene at one point in time, change positions and perform some actions but we are interested only in those who perform the given handball action.

To calculate the activity measure (A_b), for each detected bounding box (B), represented by its lower-left corner (x_B, y_B) and width w_B and height h_B in each frame, the average optical flow magnitude within that box is used:

$$A_b = \frac{1}{w_B h_b} \sum_{x=x_B}^{x_B+w_B} \sum_{y=y_B}^{y_B+h_B} |V_{x,y}| \qquad (1)$$

To filter active from inactive players an activity threshold can be used. As an alternative, in cases when only one player is active at a time in the video, i.e. only one player is performing an action, the player with max activity measure can be selected, Fig. 5.

Fig. 5. Detected players (yellow box) and detected active player (white box). IDs and activity measures are shown above the boxes.

C. *Active player tracking*

The Mask R-CNN works on a single video frame so there is no information on the boundary boxes that relate to the same objects in the successive frames. Therefore, to track players

across the video and obtain their trajectories an extra step of post-processing is required.

The track ID is assigned to each bounding box in the first frame with a confidence value greater than the threshold. Then, for each subsequent frame, assignment of the detected bounding boxes to track is done based on the Munkres version of the Hungarian algorithm [13], which minimizes the overall cost of assignment of each detection to the track. The assignment cost of adding a particular detected bounding box on each track depends on the location of the bounding box with respect to the path and the difference between the box size and the last known box in the track.

More formally, a new bounding box to be added to the track is one that corresponds to the minimal cost computed as a weighted sum of the Euclidean distance between the detected bounding box centroids (C_b) and the predicted track centroids ($C_{b+1}{}'$) taking into account the absolute difference in the area of the detected boxes (P_b) and the last box assigned to the track (P_{b-1}) (1):

$$(C_b, P_b) = \underset{b}{\operatorname{argmin}}\, w \sum_{b \in B} (d_2(C_b, C_{b+1}{}') + |P_b - P_{b-1}|); w \in [0,1];\ d_2(C_b, C_{b+1}{}') < \mathrm{T};\ \mathrm{T}, B \in \mathbb{N} \quad (2)$$

A minimal distance between locations of the bounding boxes in consecutive frames is used to predict the location of the bounding box in the next frame. This simple assumption to choose the closest bounding box in consecutive frame performed rather good, especially because a full frame rate of source video was considered and the players in most cases do not drastically change their positions even though they move at some moments quite quickly.

The cost threshold T is used to set the maximum allowed distance between the detected bounding box and the track. A box with the cost of adding to a track higher than that threshold cannot be assigned to a track, although it may be the closest to the track. This restriction has been set to mitigate detection errors and limitations of the camera view field since players may at any time enter or exit the field observed by the camera. If there are no detections for the M consecutive frames, the new detection will not be added to the track. Thus, the number of tracks can change throughout the video, and some tracks should resume after a period where no detection has been assigned. Values M and T are experimentally set to 20 and 100.

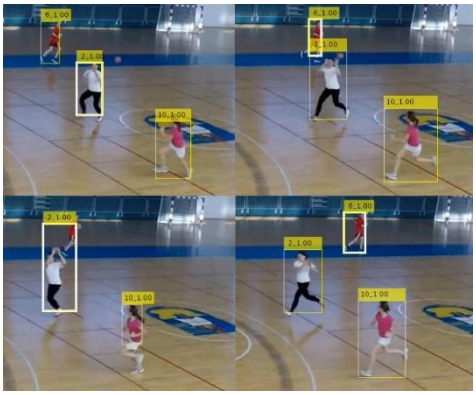

Fig. 6. In each video frame, a bounding box has its track ID

Fig. 6 shows some examples of bounding box identification with ID and tracking through different frames. The bounding boxes with IDs 2 and 6 are marked with thick white bounds in different frames, meaning that at different times different players have the highest activity measures.

To choose one player as active throughout the sequence, an additional step is performed. The detected bounding boxes are tracked across the whole sequence to form player trajectories, the Active player score is then calculated as the average activity measure of the player along the trajectory. The result is a set of player trajectories with corresponding player activity scores (Fig.7).

Fig. 7. Detected leading player (white box) and his trajectory through the whole sequence (yellow line)

D. Active player action collage and labeling

The result of the proposed method is the action collage that corresponds to the trajectory of the active player and contains all phases of an action captured in a sequence of bounding boxes along with the action label. Labeling of the college is done manually using labels from the predefined dictionary of handball techniques and actions.

In Fig. 8. an example of action college is shown for jump-shot action. The collage presents all phases of jump-shot action from running, take-off, flight, throw, and landing.

Fig. 8. Active player collage for jump-shoot action.

IV. Implementation and Discussion

We have tested the proposed method for the task of building a custom handball dataset for action recognition. The starting point was practice and competition footage acquired during a handball school. The filming was done using consumer-grade equipment (stationary GoPro Hero 4 and Hero Session 4 cameras) from different angles which represents a realistic amateur setting. The cameras were mounted at a height of around 3.5 m to the left or right side of the field for indoor scenes, while for the outdoor scenes, the cameras were at a height of 1.5 m. Full HD resolution (1920x1080) at 30 frames per second was used for all recordings.

The raw footage was first manually cut so that each video contains the whole performance of one handball technique of interest. The resulting set consisted of 600 videos containing performances of 4 actions: dribbling, passing, shooting and jump-shoot. On average, about 10 players appear in each video, and each of them can perform an action. Each file is manually labeled with a single "main" action of interest that is performed

by one or more players at the same time. The labeled actions are usually performed by one leading player in cases of shooting, jump-shooting, and dribbling actions, while two players are always involved in the case of passing the ball action. The actions other players may perform, such as running are not labelled. The total duration of labeled action videos is 1690s.

When testing the proposed MOF method, individual steps are separately tested, such as player detection, active player computation, and determination of collage action.

Performance of the Mask R-CNN detector was tested on the dataset for player detection [14]. The detector performance was evaluated in terms of recall, precision, and F1 score [15] that are shown in Fig. 9. Detection was considered a true positive when the intersection of the detected bounding box and the ground truth box was above the threshold set at 0.5. Experiments have shown that the performance of the Mask R-CNN detector depends on the number and size of the players on the scene, the contrast between a player and a background, illumination, etc. but nevertheless the results were good enough for the first step in building a learning dataset.

Player detector/Measure	*Accuracy*	*Recall*	*Precision*	*F1*
Mask R-CNN	76%	76%	98%	85%

Fig. 9. Mask R-CNN results of player detection

For evaluation of the active player detection, we have considered only the cases when there is only one leading player in the sequences. Active player detections were considered true positive only when the active player is correctly detected throughout the whole sequence, and when the trajectory of the active player is set well. This measure is determined quite strictly since the correct detection of the active player trajectory is important for correct collection of action phases and for preparation of well-defined examples of actions to be stored in the database. The true positive rate is calculated as the number of true positives divided by the number of tested sequences. The average true positive rate achieved for four handball actions is 45,96% with a significant difference between the results of detection of a particular action, from 40% for jump-shoot to 59% for passing the ball.

In more than half of the cases, a player is mistakenly selected as active, while players are properly detected and tracked. This case could be straightforwardly corrected by properly setting the appropriate active player ID. In the cases where the true active player was not properly tracked the correction could not be done easily because the collage did not capture the appropriate stages of an action and couldn't be used as action example. The solution is to manually mark action or to exclude it from the database.

Action collage that includes phases of a particular action and action name is stored in a learning set for supervised machine learning of action classifier. The aim is to generate as many consistent and well-defined examples of actions as possible, so each collage is analyzed before storing into learning set.

Each player performs a given technical element according to the rules of the game, but in a specific way with respect to his morphological and motor skills, Fig. 10. It can be seen that the collage can contain a different number of action thumbnails so that the same action phase is displayed in several frames. The number of thumbnails depends on the player activity during tracking and on the performance of the detector in all stages of the action, but the repetition of some of the action phases in a collage does not affect the correctness of the examples in action learning set.

Fig. 10. Examples of shooting action collages

Figure 11 shows examples of correct passing action collages but with occlusion of the active player. It's about mapping actions on real-life handball training sessions where players overlapping is common because of a large number of players participating at the same time in training or playing.

Fig. 11. Examples of occlusion in action collages

To make the collage action a good example of an action, it is not enough for a player to be detected, but it is important that his whole body is included in the bounding box to store the appropriate action phases in the collage. The problem of imprecise detection can be seen in Fig. 12, and should be reduced by additional detector training with the handball domain examples.

Fig. 12. Examples of imprecise player detection

All players are detected and tracked in the video and for those with an activity level above the threshold, the collage is defined. But in some cases, players who did not perform the default action had a higher level of activity than defined by the threshold, and in this case, the collages were created containing actions such as preparation for the action or coming to the queue, Fig. 13.

Fig. 13. Examples of actions outside of the scope

The errors may occur when the players are not properly detected or tracked across the video. The problem caused by poor player detection is shown in Fig. 14 and the problem of

incorrect player tacking due to cluttered scenes, overlapping or changing the player's positions in Fig 15.

Fig. 14. Example of incorrect shooting action collages due to poor detection

Fig. 15. Examples of problems in player tracking

Further training of Mask R-CNN detector on handball examples should contribute to more precise detection of players in different phases of action, which will indirectly contribute to a more successful collection of action collages.

V. Conclusion

Large datasets of labeled training examples are required to build successful machine learning models. To facilitate the creation of such dataset for handball action recognition, we present an approach for detection and marking of handball actions in sports videos, so that the amount of tiresome and time-consuming manual labor required to label the individual players performing the chosen actions is minimized.

Existing deep learning object recognition methods are employed to automatically detect the players, and the obtained location information is combined with a player activity measure based on optical flow estimate to detect the players that are performing the currently relevant action, i.e. the active players.

We have used the proposed MOF method with a challenging dataset of real-world handball practice videos, where multiple players are simultaneously present on the scene and can perform different actions. The proposed method has significantly simplified and sped-up the process of the action dataset building, as only the action beginning and end should be manually selected on the video, while the players were detected and tracked automatically. Extracted action collages with thumbnails of action phases were formed and stored in the learning dataset. Well-defined action collage containing all correct thumbnails of action phases was defined in 46% of cases. In rest of the cases, most of the time (85%) the active player was properly detected and tracked but another player was mistakenly selected as active. Here, to correct the dataset entry, only the correct active player ID should be selected manually.

In order to improve the rate of automatically obtained well-defined action collages, we plan to refine [16] and extend the method for calculating the player activity. In addition, we plan to train the Mask R-CNN detector on additional examples from the sports domain to improve the performance of player detection and precision in selection of action phases. Also, to automatically handle changing player activity during longer video sequences and to reduce requirements for manual segmentation, the method for activity recognition and computation should be extended.

ACKNOWLEDGMENT

This research was fully supported by Croatian Science Foundation under the project IP-2016-06-8345 “Automatic recognition of actions and activities in multimedia content from the sports domain” (RAASS).

References

[1] Lin, T, et al. (2014, September). Microsoft coco: Common objects in context. In European conference on computer vision (pp. 740-755). Springer, Cham.

[2] Schuldt, C., Laptev, I., & Caputo, B. (2004). Recognizing human actions: a local SVM approach. In Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on (Vol. 3, pp. 32-36). IEEE.

[3] Blank, M., et al. (2005, October). Actions as space-time shapes. In Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on (Vol. 2, pp. 1395-1402). IEEE.

[4] Ivasic-Kos, M; Iosifidis, A; Tefas, A; Pitas, I. Person De-identification in Activity Videos; BiForD - Biometrics & Forensics & De- identification and Privacy Protection, Rijeka : MIPRO, 2014. 75-80.

[5] Vondrick, C., Patterson, D., & Ramanan, D. (2013). Efficiently scaling up crowdsourced video annotation. International Journal of Computer Vision, 101(1), 184-204.

[6] Yuen, J., Russell, B., Liu, C., & Torralba, A. (2009, September). Labelme video: Building a video database with human annotations. In Computer Vision, 2009 IEEE 12th International Conference on (pp. 1451-1458). IEEE.

[7] All, K., Hasler, D., & Fleuret, F. (2011, June). FlowBoost—Appearance learning from sparsely annotated video. In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on (pp. 1433-1440). IEEE.

[8] de Souza, C. R., Gaidon, A., Cabon, Y., & Pena, A. L. (2017, July). Procedural generation of videos to train deep action recognition networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Vol. 2).

[9] K. He, G. Gkioxari, P. Dollár and R. Girshick, "Mask R-CNN," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017, pp. 2980-2988.

[10] Royden, C. S., Holloway, M. A. (2014). Detecting moving objects in an optic flow field using direction-and speed-tuned operators. Vision research, 98, 14-25.

[11] Barron, J. L., Fleet, D. J., & Beauchemin, S. S. (1994). Performance of optical flow techniques. International journal of computer vision, 12(1), 43-77.

[12] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, June 1, 2017.

[13] Munkres, J. (1957). Algorithms for the assignment and transportation problems. Journal of the society for industrial and applied mathematics, 5(1), 32-38.

[14] Burić, M. Pobar, M. Ivašić-Kos, M. "Object Detection in Sports Videos," *2018 MIPRO*, Opatija, 2018.

[15] Ivasic-Kos, M. Ipsic, I., Ribaric, S. A knowledge-based multi-layered image annotation system. Expert systems with applications. 42 (2015), 2015; 9539-9553.

[16] Ivasic-Kos, M; Pobar, M; Ribaric, S. Automatic image annotation refinement using fuzzy inference algorithms, European Centre for Soft Computing, IFSA- EUSFLAT 2015, Gijón, Asturias (Spain) : IFSA-EUSFLAT2015, 2015