The researchers in this paper created the experiment due to the increased popularity of 1st person POV action cameras used for sports as well as the resulting footage, but a lack of an effortless way to both categorize the type of footage being taken as well as the type of 1st person POV action categories (called ego-actions categories). This research stood as a proof-of-concept for ego-action learning and to how a stacked Dirichlet Process Mixture (DPM) model would automatically learn from motion histograms and set categories. In doing so, this could provide convenience for humans in the future to quickly categorize and sort through POV sports video.

The researcher used the method of unsupervised learning, no labeled training videos, and the videos are not pre-segmented, and the number of ego-action categories are unknown. The input videos consisted of a choreographed video (QUAD) with 124 video splices and 11 different ego action categories, a second 25-minute workout choreographed video (PARK) of 766 video splices and 29 categories, along with 6 public YouTube videos each of different sporting categories. The videos were of biking, surfing, skiing, slope style, snowboarding, and horseback riding. All videos were shot with a GoPro HD and the videos were processed through the models on an unspecified 2.66Ghz CPU.

The types of models included in the study were Dirichlet Process Mixture (DPM) with Online Inference (DPM-OL), DPM with Variational Inference (VI) (DPM-VI), Latent Dirichlet Allocation with VI (LDA-VI), Probabilistic latent semantic analysis with EM algorithm (PLSA-EM), naïve Bayesian mixture with EM (NBM-EM), and K-means clustering. Before video was inputted into these models, they were first converted into a 52-dimension motion histogram consisting of 36 bins for directional action data and 16 bins for periodic (frequency) action data. The directional component consisted of 4 flow directions, 3 flow magnitudes, and 3 flow variances. The histograms were then input into the various models repeatedly for higher accuracy, and the outputs were more histograms consisting of ego-actions for each video splice that were discovered because of the procedure.

The end results were scaled using an F-score with range from 0 to 1 (higher is more accurate) to determine testing accuracy. Skiing, surfing, and snowboarding all averaged above 0.6 (range 0.64-0.97) while horseback riding, mountain biking, and slope style averaged below 0.6 (range 0.47-0.54). And issue discussed regarding the variation between results was that the first three sports had stronger periodic actions while the last three did not. Additionally, the last three sports had actions and subjects that were closer to camera and larger in the frame, which resulted in greater difficulty in calculation. The QUAD video ended up with an F-score of 0.93 when using DPM-OL and PARK had an F-Score of 0.71 and superior performance here was attributed to the videos being choreographed as well as having more periodic actions. DPM-OL was superior to the other models briefly discussed prior, although the other models were generally still within 20% of the DPM-OL F-score. Something to note was that due to the length of the workout video, the athlete in the video become more tired as the video went on, which meant that some actions became harder to discern. The current model was not able to account for these changes.

This research was a novel idea because there was not prior research done within the field that was like this. The closest related works consisted of POV recognition of hand movements in an indoor environment, and other works would only have either POV object recognition, hand tracking, and these studies were often supplemented with body sensors to provide better data, unlike this experiment that only used video data. Additionally, some of the related research also used supervised learning (human aided) rather than unsupervised learning (mostly computer number crunching) which ended up costing more human labor and took more time. With the various models used in the experiment, the videos were processed and outputted between ranges of 0.25s-12.12s for QUAD, and 3.75s-73.64s for PARK (recall that this is a 25-minute video) which on time alone, a human would not be able to get even close to those speeds when even if watching the videos on higher speeds.

On a technical aspect, the paper looked mostly correct, although when discussing the results, there was a place where F-score for PARK was stated to be 0.72, but the in the following table F-score was 0.71. Whether this was due to rounding or just a typo is unknown, but a research paper of this level should be proofread for consistency as different results being stated could result in questioning the validity of the whole experiment. Otherwise, there are some clarity issues since the QUAD video was not described outside of what ego-action categories were discovered and put into a table, so it is unknown what the video was about. The technical code, algorithm, and model portions were very detailed in terms of explaining what was done to get from input to output, but this was hard to check for correctness due to the high difficulty of the subjects. Additionally, while in the beginning the researchers discussed a decent bit about how POV cameras had lots of distortion due to the environment and camera capabilities, there was not much that was used to account for all those distortions other than using more general directions and magnitudes (flow vectors) when running the models.

Finally, the video sample size was relatively small, and it would have helped to increase the variety in terms of sports categories, as well as increase number of videos for each category to determine if the models worked well for other videos. This would introduce more complexity when determining results, but having a single video for each sport seems scant, and with how fast the videos were processed, it should not take that much longer to process some more videos. On the other hand, different videos would also have different environments, people, and how they perform ego-actions, so there would be discrepancies even within each sport category, although this could further affirm how well the models can work for a variety of video inputs. Overall the research paper had good originality in terms of study being done and the methods were described in very fine detail. There would be more factors to consider when performing the study as well as more clarity in some parts of the paper, but the paper was generally fairly clear and concise.

Citation

Fast Unsupervised Ego-Action Learning for First-person Sports Videos. Kris M. Kitani, Takahiro Okabe, Yoichi Sato, and Akihiro Sugimoto. CVPR 2011 [[pdf](http://www.cs.cmu.edu/%7Ekkitani/pdf/KOSS-CVPR11.pdf)]