# Change-Point Detection Methods for Body-Worn Video

## Original Paper Authors:

Stephanie Allen, David Madras, Ye Ye and Greg Zanotti

## Original Paper Link:

<https://arxiv.org/pdf/1610.06453.pdf>

## Summary of Paper:

This paper proposes a model of processing and labeling large datasets of body-worn video, using various machine learning techniques. The paper proposes a two-stage framework, making use of computer vision methods, machine learning models and change point detection algorithms.

The first step uses two mutually exclusive, collectively exhaustive states, positive and negative, to build a compact representation of each frame. Using Scale Invariant Feature Transform (SIFT) the model is invariant to image scaling and rotation. After extracting SIFT features from the frames, 20% of the frames in the training set are taken to the Bag of Visual Words (BoVW) a form of Vector Quantisation (VQ), however these methods do not include spatial information. Using Pyramid Match Kernel (PMK) and Support Vector Machines (SVM) spatial data is interpreted. This generates a time series of frames, which are then labeled using a pre-trained deep neural network, the VGG-16, constructing a time series of scores. The time series of scores is converted into binary (0,1).

The second stage uses change-point detection algorithms to analyse the scores or labels, this identifies salient changes between the two predefined states. The model uses Mean Squared Error Minimization (MSE) to determine sequences with a single change point, then this method is extended to multiple change points. Forecasting methods are used to fit a model to a set of data then predict future observations, they develop what they have called “future window technique”, where they make a baseline model and assume that no change will occur within the first few observations of the time series then make prediction based off the baseline. The Hidden Markov Model is applied, with the ground truths of states play the role of a sequence of discrete states. Which is then has the Maximum Likelihood Estimation applied to find the most likely outcome of the data.

## Relevance of Paper to our Project:

The paper provides a model which has the application on large datasets, this is a necessity for the application in our project, as we require speed of processing for users. However, this model is heavily supervised, for our use it would need to be unsupervised. The application of this method to hand gestures/actions seems like a fairly good fit.

# ModDrop: Adaptive Mulit-Modal Gesture Detection

## Original Paper Authors:

Natalia Neverova, Christian Wolf, Graham Taylor and Florian Nebout

## Original Paper Link:

## <https://arxiv.org/abs/1501.00102>

## Summary of Paper:

This paper proposes gesture detection using a deep learning based multi-modal and multi-scale framework with localization and recognition with the capabilities of being augmented with channels of arbitrary nature. Using a tree structured deep learning architecture allowing to classify hand gestures with a higher accuracy while restricting the free parameters.

The paper formulates a pose descriptor by normalizing joint positions as well as velocity and acceleration then augmenting the descriptor with characteristic angles and pairwise distances. The skeleton is represented as a tree structure, then from the top down iteratively normalize each skeleton segment. Gaussian smoothing is applied along the temporal dimension to decrease the influence of skeleton jitter. Pairwise distances are then introduced between all joint positions. Right hand and left hand joints are then cropped to eliminate a person position with respect to the camera. Within each set of frames forming a dynamic pose, hand positions are stabilized by minimizing inter-frame square-root distances. Inter-modality fusion is used on a pre-trained data set to initialize shared layers.

The pre-processed data has Multi-modal dropout applied to it, to learn a shared model while preserving uniqueness of per channel features and avoiding false co-adaptions between modalities. The key idea is to train the model so that it would produce meaningful predictions from an arbitrary number of available modalities. Inter-scale fusion is applied during the test to obtain per class network outputs via per-frame aggregation and temporal filtering. Finally a simple localization is used to determine resting moments from periods of activity.

## Relevance of Paper to our Project:

This paper has some relevant content with regards to gesture detection however, the manner they go about processing this is a larger scale than what we’re looking at. If this process could be scaled down to forearms and hands, this process would apply well to our project. Due to the pre-training and localization the model appears to perform well in real time.