**Article 1: Zisserman A, Simonyan K (2014). Two-Stream Convolutional Networks for Action Recognition in Videos. Retrieved from** <https://arxiv.org/pdf/1406.2199.pdf>

This report investigates a process that could be implemented to the project that improves the accuracy of defining an action when given a video to be processed by a neural network. The method that they investigate consists of generating an optical flow and then computing it with the video and then averaging the outputs of both or using a Support Vector Machine to increase the accuracy slightly more.

An optical flow consists of various images that are computed from the original video source that shows the differences between frames. The way that differences are shown is done in 3 different ways, the first way adds a box on each frame around the area that contains movement between itself and the next frame. The second way is a closeup of the movement that is converted to arrows that show the difference between itself and the next frame and the last is two black and white heat maps that consists of the difference between frames being highlighted with the first one using white on black and the second black on white.

Due to using this process instead of the video data on its own the accuracy has increased from 73% with just the video data to 86.9% when averaging the result of both processes. It can even go higher when using a Support Vector Machine to generate the result instead of averaging with an accuracy of 88%.

The way that this report could be useful towards the project is that it gives us an idea on what ways we could pre-process the data before putting it into a neural network to improve the accuracy. It shows that Instead of inputting them as just images in a sequence we could also generate an optical flow that could also be computed in a separate neural network and then get the averages of it and the sequence of images to obtain a more accurate result.

The main difficulty that I think we will come across when adopting this process over a standard image recognition implementation would be trying to figure out how to generate the optical flow from the original video.

Once the image generation is figured out it should be somewhat straightforward to implement since it should be able to run parallel to the main neural network. This means that no modification of the main network should be needed and all that would be needed is a second network with the optical flow as input instead of the video. Then the final modification would be averaging the results from both processes instead of just outputting the result from the main network. Due to this difficulty level compared to the accuracy increase I think it will make a useful addition to the recognition process that we implement.

**Article 2: Molchanov P, Yang X, Gupta S, Kim K, Tyree S, Kautz S (2016). Online Detection and Classification of Dynamic Hand Gestures with Recurrent 3D Convolutional Neural Networks. Retrieved from** <http://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Molchanov_Online_Detection_and_CVPR_2016_paper.pdf>

This report considers a way to classify hand gestures using a 3d convolutional neural network. The way that they approach the issue is by breaking the video up into clips of 8 frames and computing them into a 3d convolutional neural network. The result of the convolutional neural network is then passed into a relational neural network that contains a hidden state computed from the previous clips with the output being the new hidden state. The new state is then inputted into a SoftMax layer and after all clips have been computed the final classification is then generated by averaging the results of all the clips.

The report also goes into types of data that can be imputed into the neural network to increase the accuracy of the output. It consists of using, Infrared sensors, depth sensors and an optical flow to get a more accurate result. The implementation of optical flow though is somewhat different to the one discussed in the previous report in which it only focuses on using the heat map style of optical flow.

The accuracy of the reports implementation when using their own dataset versus the second-best implementation that they tested it against has an accuracy of 74% vs 69.3% when just using colour and 83.8% vs 73.4% when implementing all data types. The accuracy when using the SKIG RGBD dataset is 98.6% for their implementation vs 97.8% for MRNN.

The usefulness that this report has towards the project is that it gives us an idea on how to implement the neural network to accurately recognise gestures and actions. Even though the report used an outside in instead of inside out camera position when recording gestures the implementation should be very similar to what we could implement to recognise gestures, the only difference being the training data that is used to train our network.

Another thing that is different about the report vs our possible implementation is the use of depth and IR sensors due to the fact of it being prohibitive to a chest mounted system but the accuracy of colour plus optical flow is still very accurate.

Another way that this report is useful is that it gives us an idea on how the program could scale in the future to support mobile devices by offloading the processing of the video to a server. This is probably not to be done in the capstone timeframe though because the cost is prohibitive and the extra time to implement a mobile application and server side code is most likely out of the scope of capstone but still could be useful for anyone who wants to continue the project in the future.