# Two-Stream Convolutional Networks for Action Recognition in Videos: Literature Review 1

## About the paper:

The paper discusses the use of deep Convolutional Networks (ConvNets) for action recognition in video footage. Video decomposition is naturally split into two components, spatial and temporal. The spatial component relates to the reference frame of the video, and the specific frame being analysed. This component carries all information about the specific scene (for example a kitchen setting), and the objects in the scene. The temporal part on the other hand, focusses on the viewpoint movement, i.e. the movement of the camera, and the movement of specific objects through the scene. Simonyan and Zisserman state that ‘Our experiments on two challenging datasets show that the two recognition streams are complementary, and our deep architecture is competitive with the state of the art shallow representations’. A spatial stream ConvNet is used to recognise human action from still images. The temporal stream on the other hand, works to recognise motion through optical stream, or sets of consecutive still frames. The paper combines the two methodologies by fusing the two separate SoftMax[[1]](#footnote-1) scores through averaging of the data and through linear support vector machines. In the end the results showed that fusion significantly improved both components and outperformed the state of the art deep architectures.

## Relevance to the project:

The research in the paper is relevant to the current project us it requires the accurate detection of human action and movement, whether this is a recognised action or otherwise. The goal of the project is to be able to detect human actions, and compare and contrast how well these actions are being performed by each side of the body. Ideally, the movement detection of the system will need to be of a higher quality. Due to the fact that stroke victims may have poor movements or inability to complete some tasks, the system would also need to recognise similar patterns to those defined in the chosen deep learning algorithm. These patterns may contain noise, or contain half the action pattern as stroke victims often require therapy to improve the use of their affected side. This paper was noted due to that fact that it has a higher quality result set than that of standard analysis. However, due to the scope of the project, this in depth algorithm may be to complicated to be included.

## References:

1. Simonyan, K., Zisserman, A. (2014). Two-Stream Convolutional Networks for Action Recognition in Videos. *Nueral Information Processing Systems*, 568-576. Retrieved from <https://papers.nips.cc/paper/5353-two-stream-convolutional-networks-for-action-recognition-in-videos.pdf>
2. Stanford University. (2017). Softmax Regression. Retrieved on 20/4/2017 from <http://ufldl.stanford.edu/tutorial/supervised/SoftmaxRegression/>

# Dense Trajectories and Motion Boundary Descriptors for Action Recognition: Literature Review 2

## About the Paper:

The research paper uses trajectories to capture local motion from video footage. Features extracted through this method are those that align with the trajectories and help to characterize the shape, appearance and motion of the objects. The paper focuses on using a dense representation of data which was then analysed using motion boundary histograms to combat the effects of camera movement, and provide a strong descriptor for action recognition. To capture motion in this way, trajectories track a spatial point over time. Statistics of these trajectories are then used to discriminate different human actions. The dense optical flow (represents the absolute motion between frames) provides better quality information about the trajectories. The study uses motion boundary descriptors to account for camera motion from the video footage and remove it, leaving only scene and object motion. Spatio-temporal pyramids (STP) were used to add structure information to the analysis. The experiment was then run over 9 data sets and yielded higher results than many other trajectory methods (KLT, SIFT, dense cuboids)[[2]](#footnote-2) and outperformed the state of the art results with the addition of the STP’s. In conclusion, the report finds that motion boundary histograms used as descriptors for dense trajectories provided excellent results. The method was designed to minimise camera motion and has shown that it outperforms all current methods.

## Relevance to the Project:

Motion detection will be highly important to the project. As the activity detection shown in this paper is more for complete actions, rather than slight finger movements for example; analysis of this type would be best accompanied by the therapist. Although this method is useful to categorise recognised actions, the ability to capture all motion is relevant to stroke survivors. Due to weaknesses or loss of movement, stroke victims may have varied movement patterns when compared to that of a healthy individual. By using different optical flow algorithms and changing the descriptors in the system it can easily be tailored to recognise different action sets. The ability to characterise shape and appearance will be useful when targeting mainly arm movements in the patient’s lifestyle. Accompanied with deep learning techniques the video analysis and feature extraction proposed in the above paper would be beneficial to the project.

Wang, H., Klaser, A., Schmid, C., Liu, C. (2013). Dense Trajectories and Motion Boundary Descriptors for Action Recognition. *International Journal of Computer Vision, 103.1* (60-79).

1. SoftMax function is a mathematical regression which is a generalisation of logistic regression to the case where we want to handle multiple classes. [2] [↑](#footnote-ref-1)
2. KLT: Kanade-Lucas-Tomasi feature tracker is and approach to feature extraction.  
   SIFT: Scale Invariant Feature Transformation, used to express the displacement between key points. [↑](#footnote-ref-2)