Anomaly Detection in Video

by

Tran Thi Minh Hanh

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

The University of Leeds
School of Computing
June 2018

Declarations

The candidate confirms that the work submitted is his/her own, except where work which has formed part of a jointly authored publication has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

Some parts of the work presented in this thesis have been published in the following articles:

Hanh T. M. Tran and David C. Hogg. Anomaly Detection using a Convolutional Winner-Take-All Autoencoder. *Proceedings of the British Machine Vision Conference (BMVC), BMVA Press, September 2017.*

The candidate confirms that the above jointly-authored publications are primarily the work of the first author. The role of the second author was purely supervisory.

Abstract

Anomaly detection is an area of video analysis that has great importance in automated surveillance. Although it has been extensively studied, there has been little work on using deep convolutional neural networks to learn spatio-temporal feature representations. In this thesis we present novel approaches for learning motion features and modelling normal spatio-temporal dynamics for anomaly detection.

The contributions are divided into two main chapters. The first introduces a method that uses a convolutional autoencoder to learn motion features from foreground optical flow patches. The autoencoder is coupled with a spatial sparsity constraint, known as Winner-Take-All, to learn shift-invariant and generic flow-features. This method solves the problem of using hand-crafted feature representations in state of the art methods. Moreover, to capture variations in scale of the patterns of motion as an object moves in depth through the scene, we also divide the image plane into regions and learn a separate normality model in each region. We compare the methods with state of the art approaches on two datasets and demonstrate improved performance.

The second main chapter presents a end-to-end method that learns normal spatio-temporal dynamics from video volumes using a sequence-to-sequence encoder-decoder for prediction and reconstruction. This work is based on the intuition that the encoder-decoder learns to estimate normal sequences in a training set with low error, thus it estimates an abnormal sequence with high error. Error between the network's output and the target output is used to classify a video volume as normal or abnormal. In addition to the use of reconstruction error, we also use prediction error for anomaly detection.

We evaluate the second method on three datasets. The prediction models show comparable performance with state of the art methods. In comparison with the first proposed method, performance is improved in one dataset. Moreover, running time is significantly faster.

Acknowledgements

First of all I would like to express my utmost gratitude to my supervisor Prof. David Hogg. His guidance, feedback and encouragement have inspired me to pursue novel ideas and he have helped me to develop the research skills. I would also like to thank him and the department for funding me for various events such as a summer school and conferences during my PhD.

My sincere gratitude goes to Project 911 - Vietnam International Education Department (VIED) Scholarship which funded me for three years of my PhD. Without VIED scholarship, I would not have had the opportunity to come to UK and do my PhD at University of Leeds.

Also a big thank you to all current and ex PhD students, Aryana Tavanai, Christiana Panayi, Leo Pauly, Rebecca Stone and so many others who have discussed ideas and helped me over the last four years. My special thanks go to Duane Carey and Fouzhan Hosseini who have advised and encouraged me through the hardest times. I would also like to thank the staff in the School of Computing for a lot of fruitful discussions and support, and for providing a conductive environment to research. There are many others, too many to list here, but my gratitude goes to everyone who has helped me or been a friend to me over the years. I thank you all.

Most importantly however I would like to thank my parents and family in Vietnam. Without the never ending support and encouragement of my parents, I would not be here now. I would also like to thank my husband, Nguyen Van Quyen, who has supported me at the toughest times, understanding when I was so caught up in my work and encouraging me when things went wrong. My special thanks go to my lovely son, who has brought so much joy laughter and happiness into my PhD.

Contents

List of Figures

List of Tables

List of Notations

The following is a list of important math notations used in the thesis. In general, the following rules are used for numbers and arrays:

- Bold capital letters (e.g. $\mathbf{W}$) denote matrices.
- Bold small letters (e.g. $\mathbf{w}$) denote column vectors. A row vector is denoted by its transpose, e.g, $\mathbf{w}^T$.
- Non-bold letters (e.g. x, l, C) are for scalars.

Latin

a - Negative slope in leaky ReLU layer

$\mathbf{b}$ - Network bias

$\mathbf{b}_l$ - Network bias of layer l

C_l - The depth of output tensor of layer l

C_0 - The depth of input tensor

$\mathbf{d}$ - Feature representation of a patch

$\mathbf{E}$ - Output tensor

e - Prediction/Reconstruction error

$\mathcal{F}_t$ - A video frame at time t

$\mathbf{P}$ - Foreground patch

$\mathbf{P}_n$ - n-th foreground patch

$\hat{\mathbf{P}}$ - Estimation of the foreground patch

H_l - The height of output tensor of layer l

H_0 - The height of input tensor

N - Batch size

$\mathbf{w}$ - Decision hyperplane normal vector

r - Regularity score

s - Anomaly score

thr - A threshold for anomaly score

W_l - The width of output tensor of layer l

W_0 - The width of input tensor

$\mathbf{W}_l$ - Network weights at layer l

$\mathbf{W}$ - Network weights

(x, y, c) - The row, column and channel indices of an element in the tensor

Greek

λ - Regularization term or weight decay

α_i, β_i - Lagrangian multipliers

ξ_i - Slack variables in one-class Support Vector Machine

ν - One-class Support Vector Machine parameter

ρ - Bias

γ - Radial basis kernel function parameter

υ - Intersection over Union threshold

τ - Temporal window

$\boldsymbol{\theta}$ - A video volume

Functions

g, f - Activation function

$\sigma(x)$ - Logistic sigmoid, $\frac{1}{1+\exp(-x)}$

Φ - Feature projection function

k - Kernel function

$\mathcal{L}$ - Loss function

$||\mathbf{W}||_F^2$ - Frobenius norm of $\mathbf{W}$

$||\mathbf{W}||_2^2$ - L_2 norm or Euclidean norm of $\mathbf{W}$

$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ - Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$

$\mathbf{a} \cdot \mathbf{b} = \mathbf{a}^T\mathbf{b}$ - Dot product between column vector $\mathbf{a}$ and $\mathbf{b}$

List of Acronyms

AMHOF - Adaptive Multi-scale Histogram of Optical Flow

AE - Autoencoder

CAE - Convolutional Autoencoder

CNN - Convolutional Neural Network

Conv-WTA - Convolutional Winner-Take-All

ConvLSTM - Convolutional Long Short Time Memory

CRF - Conditional Random Field

GMM - Gaussian Mixture Model

HOF - Histogram of Optical Flow

HMM - Hidden Markov Model

KL - Kullback-Leibler

LDA - Latent Dirichlet Allocation

LSTM - Long Short Time Memory

MHOF - Multi-scale Histogram of Optical Flow

MPPCA - Mixture of Probabilistic Principal Component Analyser

MDT - Mixture of Dynamic Textures

MRF - Markov Random Field

OCSVM - One Class Support Vector Machine

PCA - Principal Component Analysis

ReLU - Rectified Linear Unit

RNN - Recurrent Neural Network