Queen's University Belfast

Enhancing Person Re-Identification using Amazon Images

Research Topic Report: Paul Lewis Johnston, 40081039

Abstract – Person Re-Identification is the ability to recognize a person across non-overlapping cameras at different times and locations. At present, person re-identification performance, using machine learning techniques, is still very low (~30-40%). In order to improve performance, one approach is to enhance the dataset used for training the system. This will be achieved using data augmentation to generate realistic synthetic images, increasing the size and variety of the training dataset. These images will be created programmatically through web scraping person images on Amazon clothing section and performing artificial background substitution with a set of background images. An existing convolutional neural network (CNN) based person re-identification system will then be trained using the new dataset and an investigation will take place to explore how changes made to the training set and deepening of the architecture effects the performance of the system, gaining insights into additional methods to then further improve the performance of the system.

1 Introduction

1.1 Person Re-Identification

Person Re-Identification is the ability to recognize a person across non-overlapping cameras at different times and locations. Accurate person re-identification facilitates the understanding of human behavior in areas covered by surveillance cameras. A direct application of re-identification is people tracking in multi-camera systems [3][4].

The Person Re-Identification problem thus arises due to the variation that occurs between the two instants the person is recorded, such as: Different poses - The person may be captured from a different angle and be in motion therefore have changed stances. Low Resolution – Typically, the quality of surveillance images is low. Illumination - The lighting, white balance and colour saturation between the two images may vary due to the use of multiple camera sources. View – Non-overlapping cameras may cause a change in the scale and perspective of the person. Background – The background may change dramatically depending on the differing locations of the cameras. Environment – Interference from crowds of people, traffic, weather etc. may hinder the clarity of the subject in the image captured.

1.2 Invariant Feature Extraction and Metric Learning

Traditionally, person re-identification involves invariant feature extraction often followed by metric learning. Invariant feature extraction is the process of measuring and deriving values for features that are invariant under common image transformations, in order to compare and classify images. Colour is often used as it has a degree of pose invariance [7][8]. However, colour is not invariant to

illumination when using multiple camera sources [8]. Texture and shape features can also be used. For example, SDALF method [10], which uses the known symmetry of human appearance to extract colour and texture features. Pictorial structures are used to isolate specific body parts in order to extract the person from the background and then measure colour and texture features [9].

Hand-designed features may not take full advantage of the information contained in the training images and are labour intensive to develop. Therefore, supervised learning approaches have been developed that distinguish between the features and variations likely to be related or unrelated to identity, such as, the Ensemble of Localized Features (ELF) approach [11], which combines multiple simple classifiers to select the features that distinguish different people. Similarly, the approach used in [12] learns to represent different body regions using different features, which are then combined to identify different people.

Once features have been extracted, metric learning uses supervised machine learning techniques [7][11][12] to compare these features using their Mahalanobis distance [13][14]. Linear Discriminant Analysis (LDA), is the simplest approach and when different constraints are additionally enforced], such as transferring the optimisation problem into the information-theoretic setting [16], it provides better performance [10. A single or multiple shot setting [17] can be used when applying metric learning, depending on the number of images of each person. Relaxed Pairwise Learning (RPLM) [15], shows that high re-identification accuracy can be achieved using simple colour and texture features with similarity and dissimilarity constraints. Metric learning and deep learning are combined by [21], which uses hand-crafted features as input to a deep-network that learns a non-linear local metric to compare images. Prior knowledge of the re-identification problem is used by [22] to cope with illumination changes and to extract low-level features, before using the features with a subspace metric learning method. It is also possible to learn verification decision function together with a distance metric to improve performance compared with a fixed verification threshold [23]. Another method that can be used to learn a distance metric is Canonical Correlation Analysis (CCA) [24]. CCA is used in conjunction with reference descriptors in [3], to achieve highly accurate re-identification given only simple features. The main drawback of the above metric learning approaches is the problem of over-fitting caused by the small size and high variability of the available re-identification datasets compared with the large number of parameters that must be learned.

1.3 Deep Learning Convolutional Neural Networks

Recently, interest in using neural networks for computer vision has been renewed. This has been encouraged by the significant performance improvements using CNNs over previously state of the art methods [18]. Learning embeddings, which involves mapping images into a low dimensional feature space, while preserving semantic relationships between the images [19], is an application of neural networks that is particularly suited to person re-identification. For example, the 'Siamese network' [19] can learn to map visually different images of the same person to similar locations in feature space, and

map images of different people to distant locations in the feature space. This requires the network to learn to discriminate between the identifying information and unimportant background variation [1].

1.4 Multi-task Learning

Multi-task learning involves training a network to complete several auxiliary tasks in addition to the main problem of interest, and has been shown to address over-fitting and improve performance, if the axillary tasks have been chosen to complement the main learning problem [5]. In the case of person re-identification, [1] uses a Siamese architecture and combines a CNN with multi-task learning, using additional attribute labelling, such as clothing, sex and pose, which do not vary under changing illumination and is shown to improve performance and help prevent over-fitting to a particular training set or camera layout.

1.5 Data Augmentation

A key to improving the performance of the algorithms is to train the deep learning system using a large, varied dataset. However, at present, person re-identification datasets are limited in size and this causes over-fitting when training a system using such data. This small set also limits the depth of the CNN that can be used, and therefore the complexity of the problem that can be modelled, since larger networks require larger sets to be trained. Methods of data augmentation, such as Background substitution and geometric transformations have been shown to increase cross-dataset performance and help reduce the problem of over-fitting [2].

1.6 Research Objectives

As the current performance of person re-identification systems is still very low (~30-40%), the main objective of this research is to investigate and explore methods of improving the performance in this area. In this project we propose to use images from Amazon clothing section to generate a large dataset of realistic synthetic images, using Data Augmentation, namely, artificial background substitution, which will then be used to enhance the training of the existing person re-identification system and reduce over-fitting. Experiments will then be carried out to investigate how various changes made to the dataset, such as, number of training images, type of clothing, type of synthetic background etc. affects the re-identification accuracy. Finally, deeper architecture will be tested with the new dataset to evaluate their adequacy to the problem.

These findings will then allow for performance improvement across the field, providing a training set that can be used to enhance other person re-identification systems and highlight the methods which have the greatest impact on the performance of person re-identification systems, encouraging further research and exploratory analysis in these areas.

2 Technical section

The proposed solution to enhance person re-identification involves using Data Augmentation to generate a large, varied dataset, which is then used to train an existing Siamese CNN with multi-task learning. The system will be tested and changes will be made to the training set and the layers of the CNN, deepening the architecture, to explore which changes have the greatest impact on performance. The proposed solution is as follows:

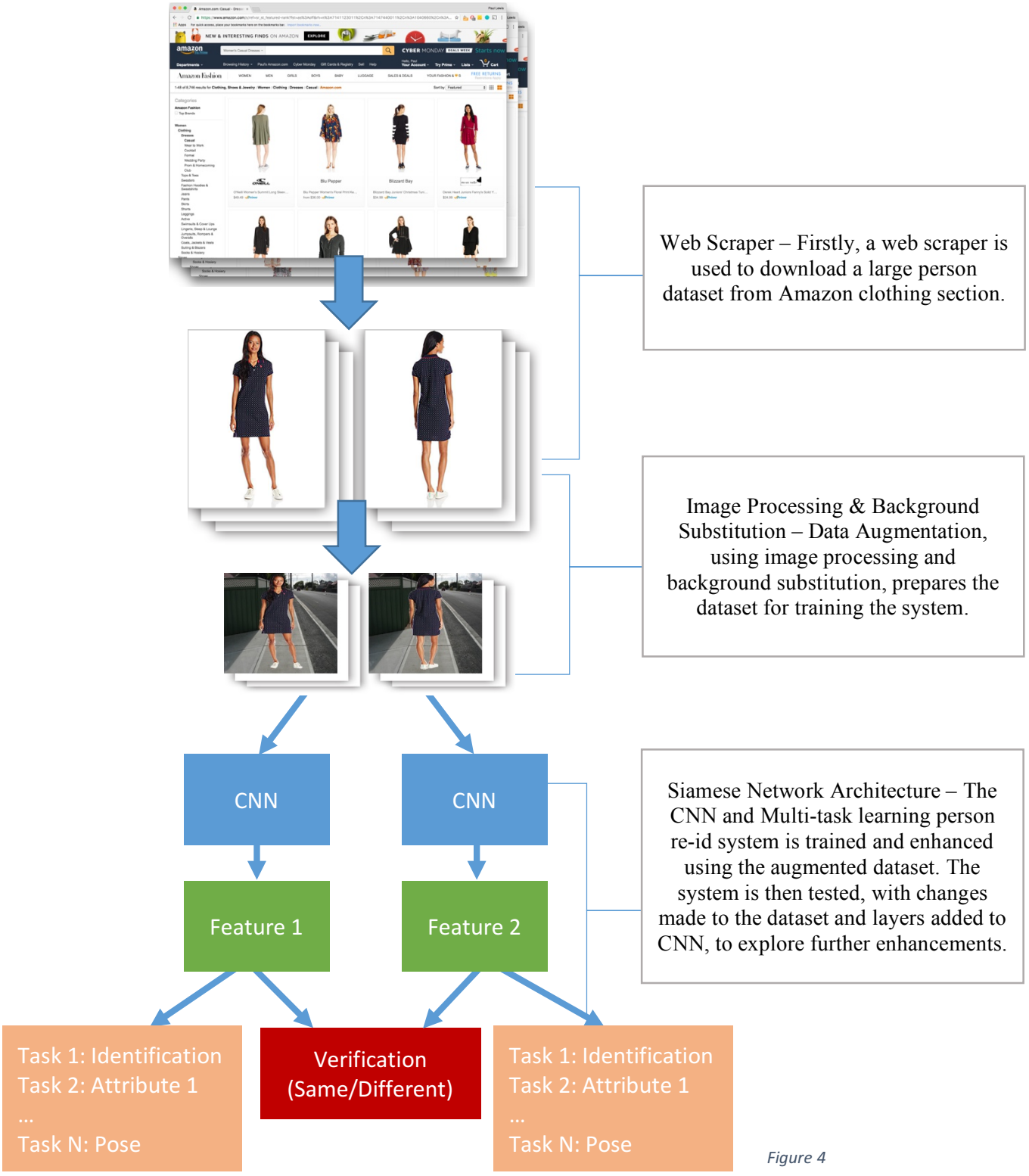

Figure 4

2.1 Data Augmentation

2.1.1 Web Scraper

The data used by the existing person re-identification system is a pair of images of the same person from different camera angles. Similar data is available in the fashion industry and we will take advantage of it. In particular, Amazon clothing section has a large database with multiple fashion brands and provides full body images, showing the person from multiple angles, standing in different poses, therefore replicating real life scenarios (Figure 2).

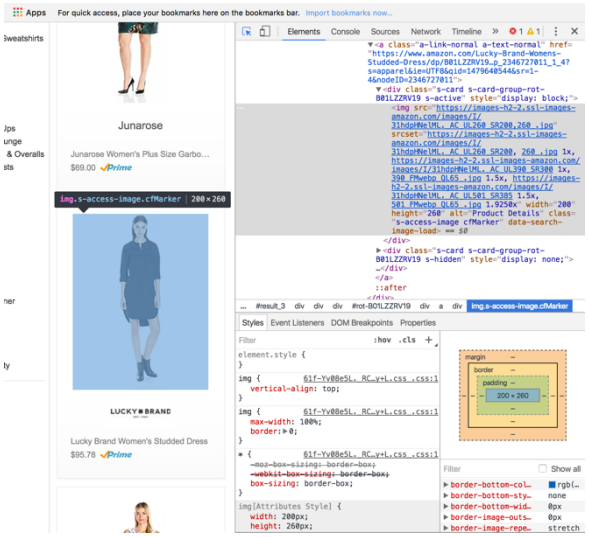

Figure 1

A web scraper will be written in Java, using an HTML web testing framework, namely HtmlUnit and/or Selenium, to programmatically download pairs of person images from Amazon clothing section. A suitable search query will be determined that returns relevant results, i.e. full body images, this may involve searching for a specific brand and/or item of clothing. Building the web scraper requires an understanding of the web pages' structure. Chrome contains a developer tool 'Inspect Element' (Figure 1) which will be used to inspect the web page and view the underlying HTML, CSS, and JavaScript in order to determine where the images are contained within the HTML of the webpage and how they can be extracted.

The web scraper will then crawl through thousands of results, locating and retrieving the URLs to the person images within the HTML. The URLs will then be used to download and save the images, using a suitable naming convention to match the pairs, thus creating a large, varied set of person image pairs. A web scraper is used to automate and speed up the process of obtaining such a large dataset from online sources. HtmlUnit is as browser-less testing framework which will be used for improved speeds

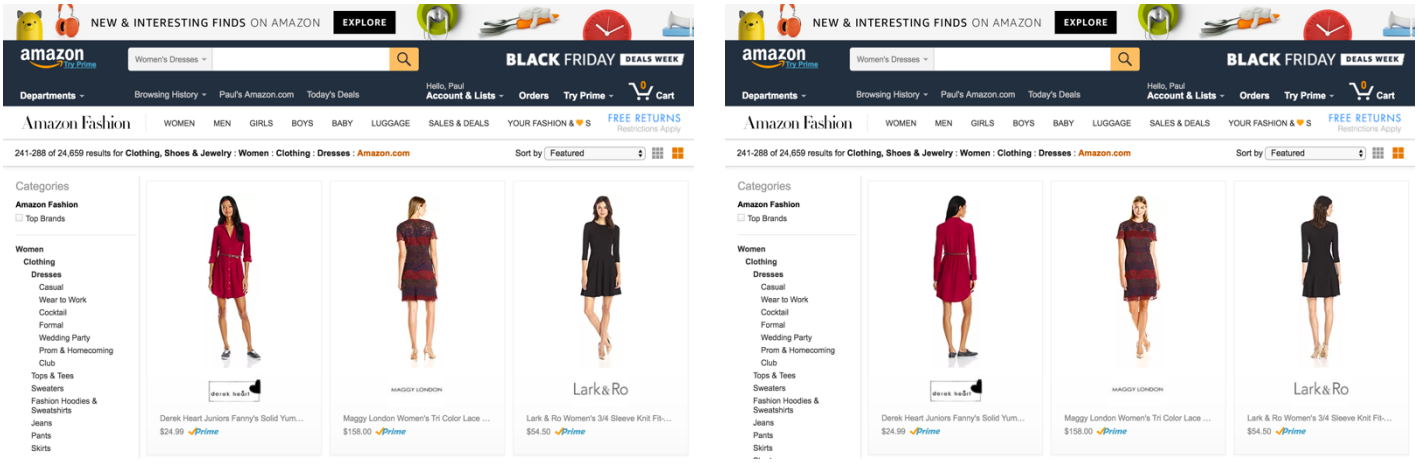

Figure 2

of web crawling. However, the additional use of Selenium may be required to replicate a browser and can be useful, if needed, to bypass any human authentication, or to invoke JavaScript functions, allowing increased access to hidden areas of the website.

2.1.2 Image Processing and Background Substitution

The background of the images on Amazon clothing section are white, this gives the additional advantage of being easy to segment. An image processing program will be written, using OpenCV Java libraries, to remove the background of the person images using various image processing techniques. This will require segmenting the image at a suitable threshold, removing the white areas, then applying erosion amongst other techniques as required. This will ensure the person is extracted effectively, removing any noise around the subject, blending sharp edges and making the background transparent throughout the dataset. The transparent background will then be substituted with an image selected at random from a set of preselected background images (Figure 3). Additional image processing techniques will then be performed to ensure the white balance, lighting and hue/saturation between the background and person of the synthetically generated images appear realistic. The images will then additionally be transformed by mirroring, cropping, rotating and changing the perspective to further replicate real life scenarios of multiple camera sources. The pairs of images will then be resized to 64x64px, as required, in order be used to train the CNN person re-identification system. Open CV Java is used as it is the most extensive open source image processing library available and the java wrapper allows for language consistency throughout the project.

Figure 3

2.2 Siamese Network Architecture

The existing person re-identification system to be enhanced incorporates the use of a deep learning convolutional neural network combined with multi-task learning. A Siamese network architecture (Figure 5) is trained to perform verification on a pair of images. A Siamese network consists of two identical sub-networks with the same network parameters that are used to produce vectors for each of the pair of images, which are low-dimensional feature representations of the respective input

images, each either in the same class or different classes. The Euclidean distance is then calculated between the images and used for training the network to verify if the pair of images belong to the same or different classes.

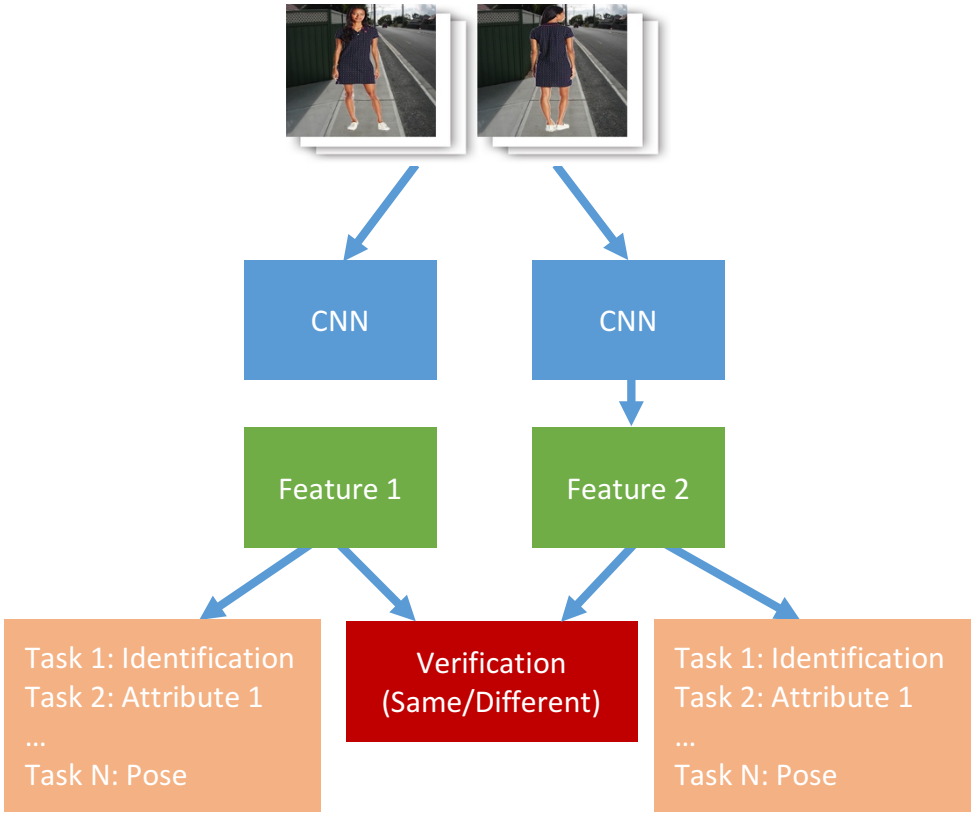

Figure 5

2.2.1 Convolutional Neural Network

The convolutional network architecture makes use of the stationary property of natural images, whereby at any given location, the statistics for the set of image patches, for a large set of natural images, are invariant [20]. This property allows sharing of network weights between image areas, significantly reducing the total number of parameters that must be learned.

Each layer of a convolutional network learns several small filters, which are convolved with the layer's input i.e., the previous layer's activation maps, to produce a new set of activation maps. Note that the filters in the first convolutional layer are connected to the colour channels of the input image. The activation maps are typically passed through a non-linear activation function, such as hyperbolic tangent, before further processing. Finally, a pooling operation, such as max-pooling [25], which takes the maximum response within a small window, is applied to the activation maps to reduce their dimensionality and to provide a small degree of translation invariance. Note that, while the network weights can be learned using back-propagation, the hyperparameters such as the number of convolutional layers, the size of the convolutional filters in each layer, and the layer widths i.e., the number of convolutional filters per-layer, are usually set by selecting the values that maximise the network's accuracy on a set of held-out validation data. The overall architecture of the convolutional network used for person re-identification in the existing approach is shown in Figure 6. This network

is composed of repeated convolutional and pooling layers, followed by a final fully connected layer that acts as the output. The hyperbolic tangent activation function was used between each convolutional layer, while a linear layer was used between the final convolutional layer and the fully connected layer. The activation of neurons in the fully connected layer gives the feature representation of the input image. Dropout regularization [20] was used between the final convolutional layer and the fully connected layer [1].

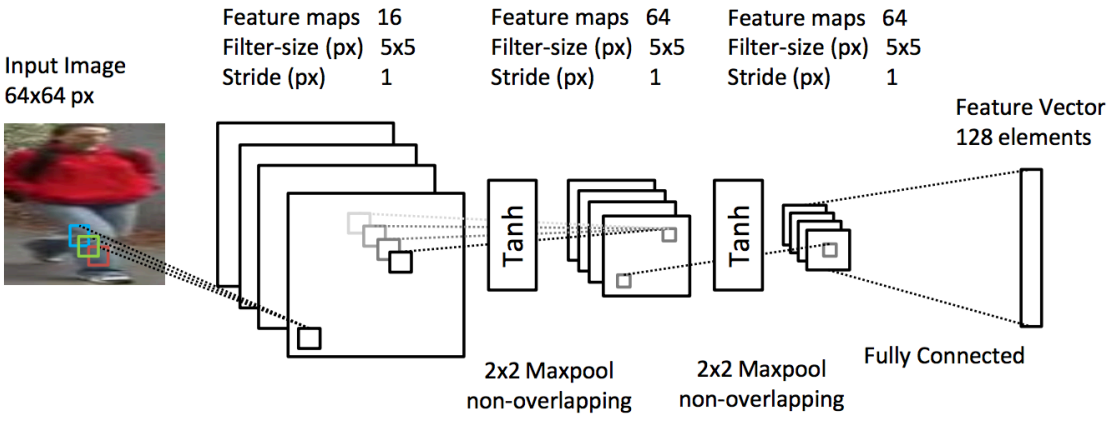

Figure 6

2.3 Validation

An experiment will then be carried out to test the performance of the existing deep learning CNN person re-identification system when the newly generated dataset is used to train the system. Firstly, the dataset prior to data augmentation will be used for training and then tested to provide a baseline of results. Then the enhanced training dataset will be used and test results will be compared against baseline. The Viper dataset will be used for testing and the performance will be measured using the cross-dataset re-identification accuracy for a better indication of real world performance [2]. An investigation will then be carried out to explore how changes made to the training set effects cross-dataset re-identification accuracy. This changes will include varying the number of training images used, combining multiple methods of data augmentation and studying underlying factors of the training set, such as, items of clothing and poses. Finally, deeper architecture will be explored to determine, suitable, additional layers that can be added to the current CNN to further improve performance of the existing re-identification system.

References –

[1] Niall McLaughlin, Jesus Martinez del Rincon, and Paul Miller. Person Re-Identification using Deep Convnets with Multi-task Learning. Journal of latex class files, vol.11, No.4, 2012.

[2] Niall McLaughlin, Jesus Martinez del Rincon, and Paul Miller. Data-Augmentation for Reducing Dataset Bias in Person Re-identification. 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). 2015

[3] L. An, M. Kafai, S. Yang, and B. Bhanu. Reference-based person reidentification. In Advanced Video and Signal Based Surveillance, pages 244–249, Aug 2013.

[4] W. Hu, M. Hu, X. Zhou, T. Tan, J. Lou, and S. Maybank. Principal axis-based correspondence between multiple cameras for people tracking. IEEE T-PAMI, 2006.

[5] G. E. Dahl, N. Jaitly, and R. Salakhutdinov. Multi-task neural networks for qsar predictions. arXiv preprint arXiv:1406.1231, 2014.

[6] D. Yi, Z. Lei, and S. Z. Li. Deep metric learning for practical person re-identification. arXiv preprint arXiv:1407.4979, 2014.

[7] R. R. Varior, G. Wang, and J. Lu. Learning invariant color features for person re-identification. arXiv preprint arXiv:1410.1035, 2014.

[8] I. Kviatkovsky, A. Adam, and E. Rivlin. Color invariants for person reidentification. Transactions on Pattern Analysis and Machine Intelligence, 35(7):1622–1634, July 2013.

[9] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. In BMVC, volume 2, page 6, 2011.

[10] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In CVPR, pages 2360–2367, June 2010.

[11] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In ECCV, volume 5302, pages 262–275. 2008.

[12] S. Bak, G. Charpiat, E. Corvee, F. Bremond, and M. Thonnat. Learning to match appearances by correlations in a covariance metric space. In ECCV, pages 806–820. 2012.

[13] K. Q. Weinberger, J. Blitzer, and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. In Advances in neural information processing systems, pages 1473–1480, 2005.

[14] W.-S. Zheng, S. Gong, and T. Xiang. Reidentification by relative distance comparison. Transactions on Pattern Analysis and Machine Intelligence, 35(3):653–668, March 2013.

[15] M. Hirzer, P. Roth, M. Kstinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In ECCV, volume 7577, pages 780–793. 2012.

[16] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Informationtheoretic metric learning. In Proceedings of the 24th international conference on Machine learning, pages 209–216. ACM, 2007.

[17] M. Hirzer, C. Beleznai, P. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In Image Analysis, volume 6688, pages 91–102. 2011.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012.

[19] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In CVPR, volume 2, pages 1735–1742. IEEE, 2006.

[20] B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? Vision research, 37(23):3311– 3325, 1997.

[21] S. Huang, J. Lu, J. Zhou, and A. K. Jain. Nonlinear local metric learning for person re-identification. arXiv preprint arXiv:1511.05169, 2015.

[22] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In Proceedings of the IEEE Conference on CVPR, pages 2197–2206, 2015.

[23] Z. Li, S. Chang, F. Liang, T. Huang, L. Cao, and J. Smith. Learning locally-adaptive decision functions for person verification. In Proceedings of the IEEE Conference on CVPR, pages 3610–3617, 2013.

[24] G. Lisanti, I. Masi, and A. Del Bimbo. Matching people across camera views using kernel canonical correlation analysis. In Proceedings of the International Conference on Distributed Smart Cameras, page 10. ACM, 2014.

[25] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun. What is the best multi-stage architecture for object recognition? In ICCV, pages 2146–2153. IEEE, 2009.