Fashion Police: Towards Semantic Indexing of Clothing Information In Surveillance Data

Owen Corrigan[1 [0000−0002−1840−982X]] and Suzanne Little[1 [0000−0003−3281−3471]]

The Insight Centre for Data Analytics, Dublin City University
owen.corrigan,suzanne.little@dcu.ie

Abstract. Indexing and retrieval of clothing based on style, similarity and colour has been extensively studied in the field of fashion with good results. However, retrieval of real-world clothing examples based on witness descriptions is of great interest in for security and law enforcement applications. Manually searching databases or CCTV footage to identify matching examples is time consuming and ineffective. Therefore we propose using machine learning to automatically index video footage based on general clothing types and evaluate the performance using existing public datasets. The challenge is that these datasets are highly sanitised with clean backgrounds and front-facing examples and are insufficient for training detectors and classifiers for real-world video footage. In this paper we highlight the deficiencies of using these datasets for security applications and propose a methodology for collecting a new dataset, as well as examining several ethical issues.

1 Introduction

In recent years(2012-2018), there have been factors to which policing organisations in western Europe have been forced to adapt. The first, and perhaps the most visible of these is a rise in the number of terrorist attacks[15][29]. Police have adapted in several ways to adapt this, including community based policing[13] and inter-agency partnerships[3]. It is difficult to say whether these initiatives have been effective, as there is a lack of evaluation research on counter-terrorism interventions[20].

Secondly, Police have also been made more effective by a variety of Information Technology(IT) initiatives. Among these is the deployment of Surveillance Video (also known as CCTV) by city councils and private companies. This can be useful for solving crimes, as police can request footage in areas where a crime may have been committed. However, CCTV data produces large amounts of data, which may require much manual reviewing.

Thirdly, there has been a dramatic increase in the performance in machine learning algorithms in recent years. Machine learning is the discipline of giving computers the ability to "learn", without having been explicitly programmed. This has been particularly successful in the domain of analysing multimedia, with tasks such as image classification[14], instance segmentation, object detection[6] and speech detection[8] becoming more accurate.

This performance improvement is both in terms of accuracy of predictions and speed of execution. These advancements have enabled new technologies to become rapidly embedded into daily life with features such as Facebook's facial recognition for face tagging, Google's photo search by concept and Apples IPhone assistant Siri. Clearly, performance has reached a threshold which has made it acceptable to be used by the general public. Along with this, there is now an awareness among the public of the impact of machine learning, with one study finding that 70% of young people in the United Kingdom "had seen or heard something about programmes which tailor web content based on browsing behaviour; voice recognition computers; facial recognition computers used in policing; and driverless vehicles"[9].

Given these three factors, a natural opportunity has arisen: to use machine learning techniques paired with the data generated from IT systems to assist police organisations to adapt to the challenge of counter-terrorism. In fact, several machine learning and statistical based products are currently deployed in other law enforcement contexts. These include an application to assess the likelihood of someone on probation from re-offending[26] and make decisions about custody arrangements[23]. There are also a number of start ups offering surveillance cameras with integrated deep learning capabilities[28].

This paper aims to bridge the gap between policing policy and machine learning practitioners, with a special focus on the application of indexing and searching videos by clothing. We have chosen a hypothetical application based on clothing search as it will be relevant to the issues which we will raise shortly.

In Section 2 we briefly review the recent advancements in the state of machine learning. While a full summary is beyond the scope of this paper, we give enough detail that a person with a background in an unrelated topic can understand the relevant capabilities. We will then describe the task of clothes classification in Section 3. This will serve as a typical example of a machine learning task in video surveillance. We will also give an overview of the publicly available datasets for surveillance and clothes detection tasks. We will highlight several deficiencies in currently existing datasets, which will motivate the discussion of gathering a new dataset in Section 3.3. We also note that that if such a system were to be deployed, it would raise some ethical questions. We outline some of these in Section 4.

2 A Brief Summary of Recent Advances In Machine Learning

In this section we will outline a very brief summary of recent advances in machine learning. As a full description of this topic is out of the scope of this paper, we recommend[16][7] for a more comprehensive view. We begin by describing four research tasks related to Object recognition in images which are of relevance, in increasing order of difficulty.

Image Classification The objective of this task is to take an image and to classify it into one of a set of predetermined labels. For example, this might

be cats, dogs, types of food, etc. This task will assign a single class for the entire image. For an example dataset, see the MNIST database[5]

Image Localisation Again, we classify an image into one of set of categories. The difference is that we also predict a bounding box in which the object appears. The previously mentioned ImageNet database contains these labels[14].

Object Detection In this task our aim is to detect multiple objects in a single image, and for each one determine the bounding box in which it is contained. This type of label can be found in the COCO dataset[17].

Instance Segmentation This task is similar to object detection, but we refine the bounding box down to pixel level segmentation. Again, these labels are found in the COCO dataset.

A visual guide to these tasks can be found in Figure 1.

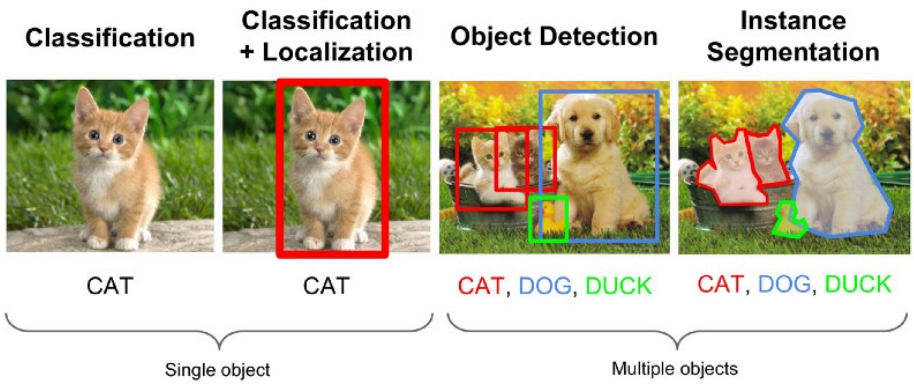

Fig. 1. Demonstration of the differences between different types of object recognition tasks. Downloaded from https://medium.com/comet-app/review-of-deep-learning-algorithms-for-object-detection-c1f3d437b852 in July 2018

In addition, we can also apply machine learning to video data sets. For example, we describe two tasks below.

Video Classification Similar to a image classification task, we take a video and determine what class it belongs to. See the Youtube8M dataset[1] for an example.

Action Recognition In this task, the goal is to predict what action is being performed in a video. For example, in the UCI101 dataset[27] some labels include "blowing candles", "High Jump" and "Cricket Shot".

Applying deep learning techniques to video techniques is a more difficult problem. Although datasets do exist for this problem they have less individual data points (e.g. videos), while having far more information to process. The state

of the art in images is more advanced, largely because there are larger data sets available for them. Even when there are large datasets for videos, they may not be as easy to label (in fact, it is harder to agree on what the labels should even be). For these reasons, in this paper we will examine of our problem under both scenarios.

We can also apply machine learning to a variety of other data sources as well, including sound recordings. However we will consider this outside the scope of this paper.

One common thread across all of the above tasks is that the state of the art performance in each of them is now held by methods based on deep learning[11][10][12][25]. This is a technique in which stacks of hidden layers (often convolutional layers in the case of image classification) are composed and trained to predict some target. This advance has been driven in recent years by a number of factors, including the availability of larger datasets, the ability to run models on a GPU and advancements in algorithms. Having briefly discussed some recent advancements in machine learning and their applications, we will now see how they can be used in a relevant practical application: clothes classification.

3 Clothes Classification Task

We will describe a clothes classification task to give a real world application of an application of deep learning. We will first consider the problem in in terms of an image recognition problem, and then extend it to video. To consider it an image problem, we extract key frames from the video and treat them as independent data points. Depending on the model of camera used, the frame rate may be anything between one image 30 seconds or a 1 frame per second.

We have seen in the previous section how object recognition can be interpreted in a number of different ways. We will now examine which of these would be the most appropriate to frame our question as. We first consider that each still from a camera is likely to contain multiple people. So object detection and instance segmentation are the most useful approaches.

Let us consider some potential applications of clothes classifier which might be useful to the police

1. Searching for a known terrorist before they flee
2. Finding lost children, based on their clothes, again given clothes description
3. Given the description of clothes of someone who has committed a crime, search through a database at a given time to find evidence of this crime

Of the above examples, none require pixel-level segmentation. So of the four image related tasks discussed in Section 2, object detection is the most useful. Similarly, object detection in videos would be the most useful application of machine learning for videos.

3.1 Feasibility Considerations

In order to determine if such an application is possible, we must determine the rate at which videos can be processed to identify clothing features. In an early deep learning paper on object localisation, RCNN, object search took 5 seconds per image[6]. In later papers this latency has been reduced down to 5 per second[24]. This has the potential to allow real time streaming of object detection. If an efficient streaming service were to be built and the frame rate of the cameras was one frame per second, we could even have one GPU service five individual cameras. This would enable a location wide search of an area with minimal hardware.

We must also consider that even if the searching does not need to be performed live, the time taken to search must still be considered. For example, if we were told that a child wearing certain clothing went missing at some point 2 days ago, we would have to search a large amount of videos. Let us say for example, the child could have reasonably walked in an area large enough to contain 50 CCTV cameras. We would like to identify the child in as quick a time as possible. If we had a single GPU based machine, at a frame processing rate of 5 per second, we would not finish an exhaustive search of the cameras for 240 hours or 10 days. This would take too long to produce actionable information. This could be reduced if a policing unit were to use more than a single GPU machine, however the number of cameras could be larger or more than one such case be active at a time. Hence 5 frames per second may not be enough. Recent results have shown even higher throughput rates[10] and have more options for trade offs between processing time and predictive performance. We expect the trend of higher frames per second processing rates for object detection to continue for the foreseeable future.

We could also imagine a computing "fog", where each camera is equipped with a GPU[4]. The idea of a "fog" computing is in contrast to "cloud" computing, where the object detection processing takes place in the camera hardware. This will be made possible as GPU inference hardware becomes cheaper. However this is a trade off which will affect the security of the system, as it is easier to access the individual cameras, and more likely that some of them will break down.

In this section we have discussed the feasibility of designing a surveillance prediction system. However, we must consider another aspect: without training data, we will not be able to build such a system. In the next section we will examine publicly available datasets.

3.2 Available Datasets

Having outlined the problem of predicting the clothes that people are wearing, and some of the reasons for which police might want to do it, we will now describe some datasets which are available for that purpose.

For Clothes Classification there are several available datasets. We highlight some of these below.

Table 1. Characteristics of Fashion Datasets vs Surveillance Datasets

	Fashion	Video Surveillance
Type	Photo	Video
Angle	Dynamic	Fixed
Source	Social Media, Retail Websites	Video Feeds
Datasets	Large, well annotated datasets available	Not many datasets available
Focus	Single Person	Can be multiple people in shot, or none
Angle	Mostly from the front	From the top
Quality	High quality	Depends on camera. Often grainy, no color.

- DeepFashion[18]. This is the most comprehensive and best labelled publicly available dataset. This dataset contains over 800,000 images. Each of these images is labelled with a category (from a choice of 50), multiple attributes (from a choice of 1000), and landmarks (e.g. left hem, right sleeve). It also contains pairs of images taken of the same item of clothing, one from social media and one taken by the shop itself. This enables us to find a piece of clothing from a shop, based on an image from social media, for example.
- Fashion MNIST[30]. This is a dataset generated by Zalando, consisting of 70,000 greyscale images of size 28×28, with a choice of 10 categories. It is mostly of interest for testing machine learning algorithms.
- Fashion 10000[19]. This paper contains 32000 images collected from Flickr across 470 fashion related categories.

We contrast this to publically available Video Surveillance Datasets.

- VIsRAT Video Dataset[22]. This dataset consists of 23 event types (for example walking, running, getting into car, entering facility) with an average of between 10 to 1500 examples per event type. This dataset is made up of ground cameras and aerial vehicles.
- 3DPes[2]. This dataset consists of hundreds of videos of 200 people taken from multiple cameras. It also contains a 3D layout of the camera coverage. The labels allow us to do perform detection, people segmentation and people tracking tasks.

We note that there are a number of differences between these two types of datasets. For example, clothes and fashion datasets typically consist of images, whereas surveillance datasets typically consist of videos. We highlight some of these differences in Table 1.

One important aspect to note is that there is far less data available in surveillance tasks than clothes tasks. This could be because it is easier to gather clothes data from the internet, and the difficultly of dealing with data privacy issues, which we shall discuss more in Section 4.1.

3.3 Proposals for Data Collection

In Section 3.2 we have examined some existing datasets for clothes classification and video surveillance. We have not been able to find a dataset which contains both videos and clothing annotations. In this section we will propose a collecting such a dataset.

This dataset should be composed of CCTV data. This could be acquired by partnering with a body such as a city council. An important choice to make initially is whether the dataset should be made public or not. Considering the nature of the collected data, it may be difficult to convince a stakeholder to release CCTV footage. However, if the dataset is made public, external researchers can compete to produce better models, which in turn improves the performance of the eventual application.

Another aspect that must be considered is whether to anonymise the faces in the dataset. An ideal dataset would have hundreds or thousands of hours of footage, with thousands of people walking in the shot. However, at this scale it would be difficult to get approval from all of them. One option might to anonymise the faces of the people in the video. However we must also remember that this may not be enoug, as their clothes may be distinctive enough to deanonymise them. We do not provide a solution to this problem, as laws and norms will vary region to region, however it is something to bear in mind.

How the data could be annotated depends on the task. We will consider some search tasks below, and then consider the annotations required. These have been selected to be useful for police agencies.

Keyword Search The aim in this task would be to get a ranked list of images of people wearing a type of clothing based on searching for a list of preselected categories. For example, a search for "heavy coats". One aspect of this would be that we would like to suppress multiple images of the same person, so that multiple shots do not pollute the search results. The data here would contain multiple bounding boxes surrounding people, each associated with a category.

Reverse Image Search The aim of this task would be to get a ranked list of images of people wearing clothes, which are similar to an image provided. An annotation scheme here might be to provide annotators with an example picture of clothes, and get them to select the image out of the dataset which most closely matches the image.

Text Search In this task, the user enters some text to describe the clothes, and a list of the ranked best matches is returned. The difference between this task and the first task is that the text here is unconstrained. For example the user could enter "a blue heavy coat with a hood" and get a list of results. Similar to the reverse image search example this could be annotated by giving some pre-selected queries and asking the annotators to select the images which best match the query.

In all of the above cases, the annotation would require a lot of person-hours. One solution to this might be to outsource the labelling required, for example to

Amazon's mechanical turk program. However, again due to data privacy issues related to this dataset, this may not be possible.

Having discussed practical issues regarding data collection and deployment of such a system, we will now discuss ethical issues related to it.

4 Ethics of Deep Learning and Surveillance Systems

In Section 1 we discussed the motivation for creating a machine learning clothes surveillance system. In Section 3.1 we have discussed the feasibility of creating such a system. We must now consider the social impact of such a system. It is important that such a system would not be deployed without public acceptance.

Of particular note, recently there has been a raised awareness of privacy issues (for example, see GDPR legislation) and the potential of machine learning algorithms to display bias[26]. In the following subsections we will outline some of these issues. We note that this is a growing field in artificial intelligence research, with topics such as accountability, transparency, explainability, bias, and fairness all topics of interest. We refer the reader to [21] for more information.

4.1 Data Protection

Managing data protection issues is a difficult topic. While collecting data, it would theoretically be possible, albeit difficult to get consent from each participant. For example, by ensuring that everyone who walks through an area is asked to complete a brief survey. The data of those who opt out could be deleted from the collected database. Alternatively, signs could be put up alerting people to the fact that data is being collected.

To deploy the system live, however, it would be impossible to get people's consent to be filmed. We could argue that they give implicit consent by walking on roads with signage alerting them to CCTV cameras, but if the sign were to placed on a road which a person has no practical alternative to crossing, the person has effectively lost their ability to consent to being recorded.

One proposal would be to have the system off in most cases, and to only enable it if there is a special event, such as an immanent threat, or a missing child.

4.2 Bias

One obvious aspect of a person's identity which is not obscured by blurring their face is their gender. Often, you will be able to tell if a person is a man or a woman just based on their clothes. Similarly, it may be possible to identify someones religion, race or age.

It is possible to imagine a situation in which a deployed system is used in a way which could lead to mis-identification of a crime suspect for wearing similar clothing. In this case the clothes a person was wearing may have been be a proxy

for gender, race or religion. In this hypothetical scenario, this system has introduced a dangerous bias. Avoiding this situation would be difficult, and perhaps the only mitigation strategy would be through training the human operator of the system.

4.3 Non Police Usage

Throughout this paper we have assumed that a deep learning surveillance system would be used responsibly by a policing organisation. We must also consider that such a system, once built, may be used by an organisation with less oversight. It could for example be produced by a commercial company, who then sell it on to shopping centres, marketing companies or directly to local councils, for example. Some examples of what might be considered unethical usage might include

- Searching areas to see if people an administrator knows personally is there currently
- Enforcing dress codes
- Sending marketing material to people based on what they are wearing

The only real protection against this is for people who have the ability to build such systems to refuse to work on them without some guarantees that the system will be deployed in an ethical manner.

5 Conclusions

In this paper we have investigated the possibility of using machine learning to detect clothes in surveillance videos. We have looked at publicly available datasets, and found that there is currently a lack of appropriate datasets to perform this task. We examined how an appropriate dataset might be collected. We also highlighted some privacy and ethical issues which might arise as a result of deploying such a system.

In future work we would like to expand on some of these issues, by collaborating with a broader range of stakeholders such as police officers and ethics researchers.

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement number 700381) project ASGARD.

The Insight Centre for Data Analytics is supported by Science Foundation Ireland under Grant Number SFI/12/RC/2289.

References

1. Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., Vijayanarasimhan, S.: Youtube-8m: A large-scale video classification benchmark. arXiv preprint arXiv:1609.08675 (2016)
2. Baltieri, D., Vezzani, R., Cucchiara, R.: 3dpes: 3d people dataset for surveillance and forensics. In: Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding. pp. 59–64. ACM (2011)
3. Boer, M.D., Hillebrand, C., Nölke, A.: Legitimacy under pressure: the european web of counter-terrorism networks. JCMS: Journal of common market studies **46**(1), 101–124 (2008)
4. Bonomi, F., Milito, R., Zhu, J., Addepalli, S.: Fog computing and its role in the internet of things. In: Proceedings of the first edition of the MCC workshop on Mobile cloud computing. pp. 13–16. ACM (2012)
5. Deng, L.: The mnist database of handwritten digit images for machine learning research [best of the web]. IEEE Signal Processing Magazine **29**(6), 141–142 (2012)
6. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 580–587 (2014)
7. Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y.: Deep learning, vol. 1. MIT press Cambridge (2016)
8. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of the 23rd international conference on Machine learning. pp. 369–376. ACM (2006)
9. Hamlyn, R., Matthews, P., Shanahan, M.: Science education tracker: Young peoples awareness and attitudes towards machine learning (February 2017), accessed: 26 July 2018
10. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Computer Vision (ICCV), 2017 IEEE International Conference on. pp. 2980–2988. IEEE (2017)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 1725–1732 (2014)
13. Klausen, J.: British counter-terrorism after 7/7: Adapting community policing to the fight against domestic terrorism. Journal of Ethnic and Migration Studies **35**(3), 403–420 (2009)
14. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
15. LaFree, G., Dugan, L.: Introducing the global terrorism database. Terrorism and Political Violence **19**(2), 181–204 (2007)
16. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553), 436 (2015)
17. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)

18. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1096–1104 (2016)
19. Loni, B., Cheung, L.Y., Riegler, M., Bozzon, A., Gottlieb, L., Larson, M.: Fashion 10000: an enriched social image dataset for fashion and clothing. In: Proceedings of the 5th ACM Multimedia Systems Conference. pp. 41–46. ACM (2014)
20. Lum, C., Kennedy, L.W., Sherley, A.: Are counter-terrorism strategies effective? the results of the campbell systematic review on counter-terrorism evaluation research. Journal of Experimental Criminology **2**(4), 489–516 (2006)
21. Meek, T., Barham, H., Beltaif, N., Kaadoor, A., Akhter, T.: Managing the ethical and risk implications of rapid advances in artificial intelligence: A literature review. In: Portland International Conference on Management of Engineering and Technology: Technology Management For Social Innovation, Proceedings. p. 682 (2016)
22. Oh, S., Hoogs, A., Perera, A., Cuntoor, N., Chen, C.C., Lee, J.T., Mukherjee, S., Aggarwal, J., Lee, H., Davis, L., et al.: A large-scale benchmark dataset for event recognition in surveillance video. In: Computer vision and pattern recognition (CVPR), 2011 IEEE conference on. pp. 3153–3160. IEEE (2011)
23. Oswald, M., Grace, J., Urwin, S., Barnes, G.C.: Algorithmic risk assessment policing models: lessons from the durham hart model and experimentalproportionality. Information & Communications Technology Law **27**(2), 223–250 (2018)
24. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. pp. 91–99 (2015)
25. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Advances in neural information processing systems. pp. 568–576 (2014)
26. Skeem, J., Eno Louden, J.: Assessment of evidence on the quality of the correctional offender management profiling for alternative sanctions (compas). Unpublished report prepared for the California Department of Corrections and Rehabilitation. Available at: https://webfiles. uci. edu/skeem/Downloads. html (2007)
27. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
28. Vincent, J.: Artificial intelligence is going to supercharge surveillance. https://www.theverge.com/2018/1/23/16907238/artificial-intelligence-surveillance-cameras-security, accessed: 26 July 2018
29. Wang, J.: Attacks in western europe. http://fingfx.thomsonreuters.com/gfx/rngs/EUROPE-ATTACKS/010042124ED/index.html, accessed: 03 July 2018
30. Xiao, H., Rasul, K., Vollgraf, R.: Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747 (2017)