MOVING OBJECT DETECTION IN VIDEOS USING PRINCIPAL COMPONENT PURSUIT AND CONVOLUTIONAL NEURAL NETWORKS

Enrique D. Tejada

Pontificia Universidad Catolica del Peru
Lima, Peru
etejada@pucp.edu.pe

*Paul A. Rodriguez**

Pontificia Universidad Catolica del Peru
Lima, Peru
prodrig@pucp.edu.pe

ABSTRACT

Object recognition in videos is one of the main challenges in computer vision. Several methods have been proposed to achieve this task, such as background subtraction, temporal differencing, optical flow among others. Since the introduction of Convolutional Neural Networks (CNN) for object detection in the Imagenet Large Scale Visual Recognition Competition (ILSVRC), its use for image detection and classification has increased, becoming the state-of-the-art in object detection and classification.

In this paper we propose to use Robust PCA (RPCA, a.k.a. Principal Component Pursuit, PCP), as a video background modeling pre-processing step, before using the Faster R-CNN model, in order to improve the overall performance of detection and classification of, specifically, the moving objects. Furthermore, we present extensive computational results that were carried out in three different platforms: A high-end server with a Tesla K40m GPU, a desktop with a Tesla K10m GPU and the embedded system Jetson TK1. Our classification results attain competitive or superior performance (F-measure) with respect to the state-of-the-art, while at the same time, reducing the classification time in all architectures by a factor raging between 4% and 25%.

Index Terms— Video Background Modeling, Principal Component Pursuit, Object Classification, Convolutional Neuronal Networks

1. INTRODUCTION

Object recognition in videos is one of the main challenges in computer vision. Recently, great advances has been made in this area thanks to the use of Convolutional Neural Networks (CNN) and Deep Learning (DL) [1], [2]. In order to correctly classify objects over the whole image, it is necessary to segment the regions to be classified. Recently, a new model has been proposed, in which the feature maps, obtained in the convolutional layers of a detection CNN, are shared with a new network scheme named Region Proposal Network (RPN) [2] to obtain this locations. This unified scheme, named Faster R-CNN, gives a state-of-the-art classification accuracy and low time response for training and classification. Other models based on CNN for object detection and classification are proposed in [3] and [4], with the drawback that they analyze the videos in batch mode.

In order to improve the overall performance of detection and classification of, specifically, the moving objects in a video sequence, we propose to use a video background modeling pre-processing step. We hypothesize that such pre-processing step, which segments the moving objects from the background, would reduce the amount of regions to be analyzed in a given frame and thus (i) improve the classification time, and (ii) reduce the error in classification for the dynamic objects present in the video. In particular, we use a fully incremental RPCA / PCP algorithm [5, 6] that is suitable for real-time or on-line processing.

Our computational results described in section 4.4 shows that the use of the PCP algorithm as a pre-processing step for moving object classification by segmenting the images with the binary mask of the sparse component, improves the accuracy of the classification.

2. RELATED WORK

Most of the work related to object detection and classification attempt to solve the problem by analyzing all the image and then determining all the objects that are present. Some of the methods used to determine these objects are based on grouping super-pixels, such as Selective Search [7], and others based on sliding windows such as [8].

Currently the Faster R-CNN model [2] achieves state-of-the-art results for object detection and classification. This model performs object segmentation via a Region Proposal Network (RPN) and, in order to reduce the computational cost, the features from the convolutional layers of the CNN are shared with the RPN. Moreover, several recent works related to object detection in images and videos, are based on Faster-RCNN model, such as DeepID-Net [9] and the solu-

*This research was supported by the "Programa Nacional de Innovación para la Competitividad y Productividad (Innóvate Perú)" Program.

tion proposed by the NUIST team in the ILSVRC challenge of 2016.

Some (task dependent) pre-processing techniques improve the classification performance, such as the method proposed in [10]. However, to the best of our knowledge, no pre-processing techniques have been previously reported for the case where the objective is to classify the moving objects in video sequences. This motived us to apply a suitable RPCA/PCP algorithm to perform a video background modeling pre-processing step and cascade it with the Faster R-CNN.

3. METHODS

3.1. Video Background Modeling via Principal Component Pursuit

In this section we give a brief overview of the RPCA / PCP method[1], with a particular focus on the incremental PCP algorithm [5, 6] (which in turn is based on [11]), which is entangled with the Faster R-CNN in order to improve the overall classification performance.

Video background modeling is a ubiquitous pre-processing step in several computer vision applications, used to detect moving objects in digital videos. There are several models for this task, e.g. based on the computation of histograms [12], subspace learning [13] and neural networks [14]. More recent models are based in PCP [15, 16] among other variants.

In particular, PCP was introduced in [15] as the non-convex optimization problem given by (1)

$$\arg\min_{L,S} \ \operatorname{rank}(L) + \lambda\|S\|_0 \ \text{ s.t. } D = L + S \ , \qquad (1)$$

where $D \in \mathbb{R}^{m\times n}$ is the observed video of n frames, each of size $m = N_r \times N_c \times N_d$ (rows, columns and depth or channels respectively), $L \in \mathbb{R}^{m\times n}$ is a low rank matrix representing the background and $S \in \mathbb{R}^{m\times n}$ is a sparse matrix representing the foreground (moving objects).

While most PCP algorithms are directly based on the convex relaxation (2)

$$\arg\min_{L,S} \|L\|_* + \lambda\|S\|_1 \ \text{ s.t. } D = L + S, \qquad (2)$$

this is not the only possible tractable problem that can be derived from (1). As it is shown in [17], (3) is also a proper convex relaxation of (1)

$$\arg\min_{L,S} \frac{1}{2}\|L+S-D\|_F^2 \quad \text{s.t. } \|S\|_1 \le \tau, \operatorname{rank}(L) \le r \ . \ (3)$$

Which can be solved iteratively via the alternating optimization

$$L_k^{(j+1)} = \arg\min_{L} \|L_k + S_k^{(j)} - D_k\|_F^2 \text{ s.t. } \operatorname{rank}(L_k) \le r \, (4)$$

$$S_k^{(j+1)} = \arg\min_{S} \|L_k^{(j+1)} + S_k - D_k\|_F^2 \text{ s.t. } \|S_k\|_1 \le \tau, (5)$$

where $L_k = [L_{k-1} \ \ l_k]$, $S_k = [S_{k-1} \ \ s_k]$ and $D_k = [D_{k-1} \ \ d_k]$. The minimization of (4) can be computed via the incremental thin SVD [18] procedure, while the minimizer of (5) is the projection of $(d_k - l_k)$ onto the ℓ_1-ball. For further details, the reader is referred to [17].

The incremental PCP algorithm [5, 6], which we use in the present work[2], exploits the particular structure of the solution proposed in [11] to transform it into an incremental one: the computationally demanding (and batch) solution of sub-problem (4) can be efficiently computed via rank-1 modifications for thin SVD (see [18] and the many references therein) that are calculated when a new (video) frame becomes available, resulting in a fully incremental algorithm that can also adapt to changes in the background (such as sudden illumination changes).

For the sake of completeness, we mention that, to the best of our knowledge, (ReProCS) [19] (GRASTA) [20], (pROST) [21], (GOSUS) [22] and the incremental PCP (incPCP) [6] are the only PCP-like methods for the video background modeling problem that are considered to be incremental. However, except for incPCP, these methods have a batch initialization/training stage as the default/recommended initial background estimate [3].

3.2. Convolutional Neural Networks

The state-of-the-art for image classification nowadays is achieved by Convolutional Neural Networks (CNN). Recently, the Faster R-CNN [2] model was presented, based on the Fast R-CNN model [23], and proposed a Region Proposal Network (RPN) for generating the region proposals, instead of the Region of Interest (RoI) pooling layer of [23].

The Faster R-CNN model has shown great performance in object classification and it has been used as a basis for new models and techniques ([9]) used by several teams in the different categories of the ILSCVR challenge, obtaining state-of-the-art results for detection and classification.

Most models are focused in detect and classify all the objects in an image, and in the case of videos, this will increase the computational cost. To solve this problem, we propose the use of PCP as a pre-processing step to perform a segmentation of the moving objects in videos and reduce the computational cost and classification time, since less regions are to be found. Faster R-CNN has shown a high performance when used along with the PCP pre-processing step.

[1] In this paper, from this point onwards, we choose to use the term "PCP"

[2] Specifically we use the variant proposed in [17], which identifies and diminishes the ghosting effect, usually observed in video background modeling.

[3] GRASTA and GOSUS can perform the initial background estimation in a non-batch fashion, however the resulting performance is not as good as when the default batch procedure is used; see [?, Section 6]. pROST is closely related to GRASTA, and it shares the same restrictions. All variants of ReProCS also use a batch initialization stage.

4. COMPUTATIONAL RESULTS

4.1. Datasets

The CDNet2014 [24] dataset was selected for the tests since it comprise several videos with particular characteristics that allow tests of moving object detection in different scenarios. We selected seven from three different categories of the CD-Net dataset

- **badWeather:** skating.
- **baseline:** highway, pedestrians, PETS2006
- **shadow:** backdoor, busStation, cubicle

4.2. Architecure

In order to assess the time performance of the proposed method[4], we have run our experiments in three different hardware platforms, labeled as "Server" (32x Intel Xeon E5-2640 CPU, 128Gb RAM, 2x NVidia Tesla K40m GPU), "Desktop" (8x Intel Core i7-2600K CPU, 32Gb RAM, 2x NVidia Tesla K10m GPU) and "Mobile" (ARM Cortex A15 CPU, 2Gb RAM, Tegra TK1). The main objective of using these different platforms was to factor out any hardware dependency in our experiments.

4.3. Procedure

For the classification of moving objects in videos, the incremental PCP via projections onto the ℓ_1-ball [17] was applied. Assuming that for any frame k, the low-rank ($\mathbf{l}$) and sparse ($\mathbf{s}$) components satisfy

$$\mathbf{d}_k \approx \mathbf{l}_k + \mathbf{s}_k, \tag{6}$$

then a binary mask $\mathbf{m}_k$ was automatically computed from $\mathbf{s}_k$. Then such mask was applied to the original frame, i.e.

$$\mathbf{u}_k = \mathbf{m}_k \odot \mathbf{d}_k, \tag{7}$$

where $\odot$ represents element-wise product.

The images $\mathbf{u}_k$ were feed to a pre-trained CNN, specifically, the Faster-RCNN [2] model with the "fast" version of ZF net [25] that has 5 convolutional layers and 3 fully-connected layers. The ZF model was chosen due to the hardware restrictions of the "Mobile" platform.

The neural network returns the bounding boxes of the images detected along with the score of classification for each bounding box, and the time needed to classify the objects in the image. This information is used along with the groundtruth for each video to determine the F-measure

$$F = \frac{2 \cdot P \cdot R}{P + R}, \; P = \frac{TP}{TP + FN}, \; R = \frac{TP}{TP + FP} \tag{8}$$

where P and R stand for precision and recall respectively, and TP, FN and FP are the number of true positive, false negative and false positive pixels, respectively.

4.4. Results

Two different test were run in each platform, first the classification was performed on the original images of the videos, and a second classification was performed on the segmented images $\mathbf{u}_k$. The F-measure was calculated for each one of the videos. In order to compute the F-measure, first we calculate the overlap ratio between the groundtruth bounding boxes and the bounding boxes provided by the Classifier using the Intersection over Union (IoU) method, this ratio allowed us to determine the metrics needed in the F-measure calculation.The performance given by the F-measure are shown in Table 2.

We first mention that, unsurprisingly, the performance results are the same for all platform. We can note that for most of the videos, the performance of the F-measure was higher. As can be seen in Table 2, the performance of the proposed method is better for most of the considered test videos after the PCP algorithm was applied.

Figure 1 (a) shows the classification of the frame 595 of the "pedestrians" video. Here we can observe that in the case of the original image, a water hydrant was classified as a person, this error was persistent through all the video, decreasing the F-measure for the video without the pre-processing step. The "cubicle" and "highway" videos are two cases for which the standard classification gave better performance. The following problems were observed while generating the masked frame ($\mathbf{u}_k$) (i) In the case of "cubicle", when the people walking by stand still for certain periods of times and the PCP algorithm considers them as part of the background as can be seen in Figure 1 (b), this problem is recurrent over all the video and thus decreases the performance; and (ii) For both videos, some of the objects lack good contrast with the background and lose some necessary features for the classification. In the case of the "highway" video, it can be noted in Figure 1 (c) that no regions were proposed for some objects although these have good contrast and have enough visible features to be classified.

The average classification time for each video is shown in Table 1. The impact on the time reduction observed when classifying the sparse images over the original images will depend on the application. It is worth to mention that the PCP time depends solely on the image size and not the content. In the case of the images classified on the server there is an improvement that ranges from 2% to 13%. For the images classified in the Desktop, a reduction from 2% to 25%. In the embedded system the reduction in the classification time ranges from 2% to 21%. More information on computational time of the PCP algorithm can be found in [26].

[4] To use PCP as a video background modeling pre-processing step, before using the Faster R-CNN model

Dataset	Server: Original frame: $\mathbf{d}_k$	Server: Masked frame: $\mathbf{u}_k$ (see (7))	Desktop: Original frame: $\mathbf{d}_k$	Desktop: Masked frame: $\mathbf{u}_k$ (see (7))	Jetson TK1: Original frame: $\mathbf{d}_k$	Jetson TK1: Masked frame: $\mathbf{u}_k$ (see (7))
backdoor	76.1	**68.7**	145.4	**123.2**	1024.1	**827.4**
busStation	81.2	**79.3**	146.2	**135.9**	1032.3	**958.9**
cubicle	82.1	**75.1**	160.4	**151.8**	1016.6	**890.6**
highway	74.5	**73.0**	178.8	**175.1**	902.5	**867.5**
pedestrians	85.0	**74.0**	224.7	**166.7**	1085.4	**858.2**
PETS2006	83.6	**77.9**	195.1	**168.4**	983.2	**861.3**
skating	80.9	**77.8**	140.5	**134.9**	919.5	**894.2**

Table 1. Average Classification times for each video tested in the CDNet Dataset, all times are in milliseconds. It can be noted that the use of the PCP algorithm for segmentation of the background objects allows a faster classification time.

Dataset	All platforms: Original frame: $\mathbf{d}_k$	All platforms: Masked frame: $\mathbf{u}_k$ (see (7))
backdoor	0.7755	**0.8309**
busStation	0.1927	**0.3801**
cubicle	**0.7505**	0.6008
highway	**0.8383**	0.8002
pedestrians	0.6094	**0.8842**
PETS2006	0.5068	**0.6231**
skating	0.4690	**0.4863**

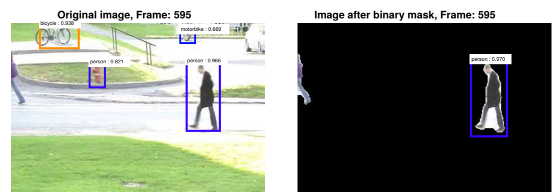
Table 2. The F-measure computed for the 7 datasets. Results are shown for classification over original frames ($\mathbf{d}_k$) and for masked frames ($\mathbf{u}_k$) (see (7))

5. DISCUSSIONS

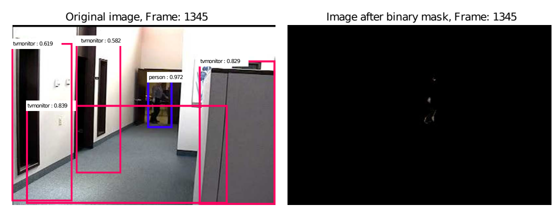
The results from Table 2 show that independently of the architecture being used, the classification performance remains unchanged as expected. One of the most remarkable results obtained is that in most of the cases the F-measure shows an improvement ranging from 3.7% to 97.2%, with a mean improvement of 22% when the sparse image was used to detect and classify the object with the neural network. The main reason for this is that the neural network finds the features of only the moving objects, instead of all the image, which can cause an increase of False Positives in classification. This can be noted in Figure 1 (a) where a water hydrant has been misclassified as a person.

6. CONCLUSIONS

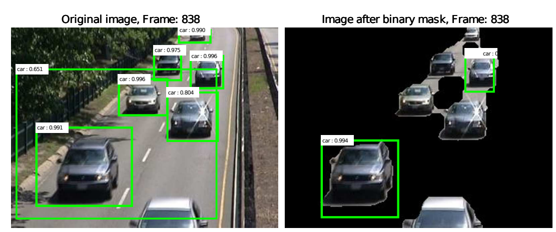
For certain applications it is important to classify the moving objects in a video, without taking care of the background. We have proved that by segmenting the moving objects with the PCP algorithm the classification performance via the F-measure increased. The classification time when the masked images ($\mathbf{u}_k$) are used show a reduction that ranges from 2% to 13% when classified in the Server, 2% to 25% when classified in the Desktop and 2% to 21% when classified in the Jetson TK1. This reduction in the classification time could potentially be beneficial for mobile applications.



(a) Classification of frame 595 of the video **pedestrians**



(b) Classification of frames 1345 of the video **cubicle**



(C) Classification of frames 838 of the video **highway**

Fig. 1. Classification samples of 2 datasets showing the benefits and drawbacks of the method proposed.

7. REFERENCES

[1] Alex Krizhevsky, Ilya Sutskever, and Hinton Geoffrey E., "ImageNet Classification with Deep Convolutional Neural Networks," *Advances in Neural Information Processing Systems 25 (NIPS2012)*, pp. 1–9, 2012.

[2] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian

Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *Nips*, pp. 1–10, 2015.

[3] Xiaowei Zhou, Can Yang, and Weichuan Yu, "Moving object detection by detecting contiguous outliers in the low-rank representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 3, pp. 597–610, 2013.

[4] Kai Kang, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang, "Object detection from video tubelets with convolutional neural networks," in *2016 IEEE CVPR*, 2016, pp. 817–825.

[5] Paul Rodriguez and Brendt Wohlberg, "A matlab implementation of a fast incremental principal component pursuit algorithm for video background modeling," *ICIP 2014*, pp. 3414–3416, 2014.

[6] Paul Rodriguez and Brendt Wohlberg, "Incremental Principal Component Pursuit for Video Background Modeling," *Journal of Mathematical Imaging and Vision*, vol. 55, no. 1, pp. 1–18, 2016.

[7] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *International Journal of Computer Vision*, vol. 104, no. 2, pp. 154–171, 2013.

[8] C. Lawrence Zitnick and Piotr Dollár, *Edge Boxes: Locating Object Proposals from Edges*, pp. 391–405, Springer International Publishing, Cham, 2014.

[9] W. Ouyang, X. Wang, X. Zeng, Shi Qiu, P. Luo, Y. Tian, H. Li, Shuo Yang, Zhe Wang, Chen-Change Loy, and X. Tang, "Deepid-net: Deformable deep convolutional neural networks for object detection," in *2015 IEEE CVPR*, June 2015, pp. 2403–2412.

[10] S. Lawrence, C. L. Giles, Ah Chung Tsoi, and A. D. Back, "Face recognition: a convolutional neural-network approach," *IEEE Transactions on Neural Networks*, vol. 8, no. 1, pp. 98–113, Jan 1997.

[11] P. Rodríguez and B. Wohlberg, "Fast principal component pursuit via alternating minimization," in *IEEE ICIP*, Sept. 2013, pp. 69–73.

[12] Yinhai; Nihan Nancy; Hallenbeck Mark Zheng Jianyang; Wang, "Extracting Roadway Background Image: Mode-Based Approach," *Transportation Research Record Journal of the Transportation Research Board*, vol. 1944, 2006.

[13] N M Oliver, B Rosario, and A P Pentland, "A Bayesian computer vision system for modeling human interactions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, 2000.

[14] A Maddalena L.; Petrosino, "A Self-Organizing Approach to Background Subtraction for Visual Surveillance Applications," *IEEE Transactions on Image Processing*, vol. 17, no. 7, 2008.

[15] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization," in *NIPS 22*, 2009, pp. 2080–2088.

[16] Emmanuel J. Candes, Xiaodong Li, Yi Ma, and John Wright, "Robust Principal Component Analysis?," *J. ACM*, vol. 58, no. 3, pp. 11:1–11:37, 2011.

[17] Paul Rodriguez and Brendt Wohlberg, "An incremental principal component pursuit algorithm via projections onto the ℓ_1-ball," in *International Congress on Electronics, Electrical Engineering and Computing,*, 2017.

[18] M. Brand, "Fast low-rank modifications of the thin singular value decomposition," *Linear Algebra and its Applications*, vol. 415, no. 1, pp. 20 – 30, 2006.

[19] H. Guo, C. Qiu, and N. Vaswani, "An online algorithm for separating sparse and low-dimensional signal sequences from their sum," *IEEE TSP*, vol. 62, no. 16, pp. 4284–4297, Aug 2014.

[20] J. He, L. Balzano, and A. Szlam, "Incremental gradient on the Grassmannian for online foreground and background separation in subsampled video," in *IEEE CVPR*, June 2012, pp. 1568–1575.

[21] F. Seidel, C. Hage, and M. Kleinsteuber, "pROST: a smoothed lp-norm robust online subspace tracking method for background subtraction in video," *Machine Vis. and Apps.*, vol. 25, no. 5, pp. 1227–1240, 2014.

[22] J. Xu, V. Ithapu, L. Mukherjee, J. Rehg, and V. Singh, "GOSUS: Grassmannian online subspace updates with structured-sparsity," in *IEEE ICCV*, Dec. 2013, pp. 3376–3383.

[23] Ross Girshick, "Fast R-CNN," *IEEE ICCV*, pp. 1440–1448, 2015.

[24] Y. Wang, P. M. Jodoin, F. Porikli, J. Konrad, Y. Benezeth, and P. Ishwar, "Cdnet 2014: An expanded change detection benchmark dataset," in *2014 IEEE CVPR Workshops*, June 2014, pp. 393–400.

[25] Matthew D Zeiler and Rob Fergus, *Visualizing and Understanding Convolutional Networks*, pp. 818–833, Springer International Publishing, Cham, 2014.

[26] Gustavo Silva and Paul Rodriguez, "Jitter invariant incremental principal component pursuit for video background modeling on the TK1," *Conference Record - Asilomar Conference on Signals, Systems and Computers*, vol. 2016-February, no. 1, pp. 1403–1407, 2016.