A New Approach Combining Trained Single-View Networks with Multi-View Constraints for Robust Multi-View Object Detection and Labelling

Yue Zhang[a], Adrian Hilton[b] and Jean-Yves Guillemaut[c]
Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, Guildford, UK
{yz01163, a.hilton, j.guillemaut}@surrey.ac.uk

Keywords: Multi-View Object Detection, Multi-View Object Labelling.

Abstract: We propose a multi-view framework for joint object detection and labelling based on pairs of images. The proposed framework extends the single-view Mask R-CNN approach to multiple views without need for additional training. Dedicated components are embedded into the framework to match objects across views by enforcing epipolar constraints, appearance feature similarity and class coherence. The multi-view extension enables the proposed framework to detect objects which would otherwise be mis-detected in a classical Mask R-CNN approach, and achieves coherent object labelling across views. By avoiding the need for additional training, the approach effectively overcomes the current shortage of multi-view datasets. The proposed framework achieves high quality results on a range of complex scenes, being able to output class, bounding box, mask and an additional label enforcing coherence across views. In the evaluation, we show qualitative and quantitative results on several challenging outdoor multi-view datasets and perform a comprehensive comparison to verify the advantages of the proposed method.

1 INTRODUCTION

Multi-view object detection and labelling is a complex problem which has attracted considerable interest in recent years and has been employed in many application domains such as surveillance and scene reconstruction (Luo et al., 2014). Compared to single-view data, multi-view data provides a richer scene representation by capturing additional cues through the different viewpoints; these can help tackle the problem of object detection and labelling more effectively by resolving some of the visual ambiguities. However, dealing with multi-view features is challenging due to large viewpoint variations, severe occlusion, varying illumination and changes in resolution (Chang and Gong, 2001).

Multi-view detection and labelling suffer from two important limitations: First, few datasets for multi-view object detection and tracking are available; Second, most approaches follow a tracking-by-detection methodology which ignores the coupling between detection and tracking. Beside, existing methods for multi-view object detection and labelling usually rely on videos, as tracking objects in both spatial and temporal domains has been shown to be beneficial due to the complementary cues they afford. However, being able to detect and consistently label objects across image views (without use of tem-

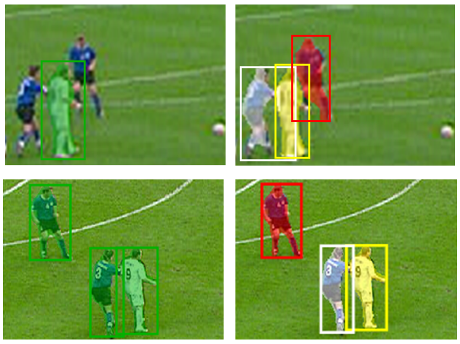

Figure 1: Illustration of the advantages of proposed approach (right) compared to the classical Mask R-CNN formulation (left) in the case of two views from the *Football* dataset. By incorporating multi-view information within the network, the proposed approach is able to reduce mis-detections while at the same time producing more consistent object labelling across views.

[a] https://orcid.org/0000-0002-3287-2474
[b] https://orcid.org/0000-0003-4223-238X
[c] https://orcid.org/0000-0001-8223-5505

Published in: Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP, ISBN 978-989-758-402-2, ISSN 2184-4321, pages 452-461. DOI: 10.5220/0008991104520461.
The final publication is available at https://doi.org/10.5220/0008991104520461

poral information) remains an important task in its own right with many practical applications. These include for example multi-view scene modelling from a hand-held camera where no temporal information is available or processing of CCTV data acquired at a low frame-rate where temporal information is too coarsely sampled. Further, multi-view image approaches are important to enable processing of keyframes in multi-view video datasets, which in turn can be used to guide multi-view video processing. It is thus essential to develop effective algorithms for processing multi-view images.

A major challenge with processing multi-view images as opposed to videos relates to the changes in object location and appearance which are usually significantly larger across views than across frames. This makes detection and tracking across views significantly more difficult in practice. Furthermore, many multi-view approaches solve the detection and tracking tasks separately, with the detection algorithm serving to initialise multi-view tracking. This sequential approach suffers from the limitation that errors at the detection stage propagate to the later stages of the pipeline, affecting tracking performance and making the approach sub-optimal.

In recent years, deep learning based approaches have achieved impressive performance in various single-view tasks such as classification, object detection and semantic segmentation (He et al., 2017; Levine et al., 2018; He et al., 2016). However, due to the lack of multi-view data, existing multi-view tracking approaches cannot achieve end-to-end deep learning. Most of the existing approaches for multi-view tasks use 3D convolution networks to train a model or classify objects in the 3D domain. However, for multi-view tasks, it is time-consuming and computationally expensive to obtain multi-view object tracking results with a 3D convolution network. On the other hand, many methods combine deep learning components separately for object detection or features trained from person re-identification.

To overcome the problem of the limited number of multi-view datasets and connect detection with labelling, we propose a new approach which integrates an existing trained single-view network with multi-view computer vision constraints. This new joint multi-view detection and tracking approach does not require further training thereby avoiding the need for annotated multi-view training datasets which are currently scarce. Our approach extends Mask R-CNN (He et al., 2017), a state-of-the-art approach for single-view image classification, detection and segmentation. The proposed framework consists of two branches, each with weights set as in the original Mask R-CNN network, which are integrated with additional components to enforce epipolar constraints, appearance similarity and class consistency thus allowing matching of instances between two views. Moreover, our multi-view framework incorporates a new branch for the label output compared with the single-view method. Figure 1 illustrates the advantages of the multi-view extension which is able to leverage multi-view information to reduce misdetections while at the same time adding label information compared to a traditional Mask R-CNN implementation.

Our approach makes the following key contributions. First, it extends a single-view deep learning network to multiple views without further training for pairs of images, introducing a framework for classification, detection, segmentation and labelling. Second, we improve the performance in multiple object detection and labelling by jointly solving these two tasks and integrating them into a common framework.

The remainder of this paper is organised as follows. In Section 2, we review related work on object detection as well as multi-view object tracking. The proposed methodology is then presented in Section 3. Section 4 experimentally evaluates the approach showing qualitative and quantitative results based on pairs of views obtained from a range of challenging multi-view datasets and comparing against established approaches. Conclusions and future work are finally discussed in Section 5.

2 RELATED WORK

Object detection: Object detection has been playing a fundamental role in a wide variety of tasks such as classification and object tracking. Thus we do not intend to conduct a thorough review here, but instead we concentrate on deep learning based techniques and joint detection methods, which are most relevant to our work. We refer the reader to recent surveys for a more comprehensive review (Zhao et al., 2018; Li et al., 2015; Hosang et al., 2016; Neelima et al., 2015). Region proposal generation is typically the first part in the object detection pipeline, where deep networks have been adopted to predict the bounding boxes and generate regions (Sermanet et al., 2013; Erhan et al., 2014; Szegedy et al., 2014). Based on those techniques, Girshick *et al.* proposed the R-CNN, which adopted an end-to-end training to classify the proposed regions (Girshick et al., 2014). To improve the detection efficiency, He *et al.* proposed the SPP-net with shared information, together with pyramid matching to correct geometric distortion (He et al., 2015). It also should be mentioned that the use of the shared information in object detection has also been applied in (Dai et al., 2015; Girshick, 2015; Ren et al., 2015), in which real-time prediction has been achieved. Based on Faster R-CNN (Ren et al., 2015), He *et al.* further developed Mask R-CNN (He et al., 2017), which achieves state-of-the-art performance for object detection and semantic segmentation. Besides, several works have made use of information from other images or views, aiming to improve the detection accuracy by sharing information among views. For examples, Xiao *et al.* built a single deep neural network for labelling re-appearing objects (Xiao et al., 2017). In this architecture, detection and identification cooperate together with shared convolutional feature maps to improve the result. In (López-Cifuentes et al., 2018), a multi-camera system is built to refine the bounding box for object detection. In their work, a graph representation is proposed by connecting different components together. Although multi-view cues can help to improve the detection, this is an area which remains relatively unexplored.

Multi-view tracking: Multi-view tracking is a broad area encompassing camera calibration, object detection, person re-identification, object tracking, etc. The existing literature on multi-view tracking, however, mostly incorporates several features and hand-crafted pipelines to combine the temporal and spatial information on multi-view videos (Luo et al., 2014). Ristani *et al.* used correlation clustering optimization to find trajectories based on the combination of appearance and motion features (Ristani and Tomasi, 2018). They then use post-processing to globally refine the result. Multi-view tracking combines several cues for representation and uses various strategies to solve the tracking problem. Xu *et al.* used a trained DCNN to represent the appearance of people, and combine geometry as well as motion to build a model. Then for each cue, a composition criterion is set and then jointly optimised to achieve correct tracking (Xu et al., 2016). Hong *et al.* also combined multiple cues into a sparse representation (Hong et al., 2013). Then they treated multi-view tracking as a sparse learning problem. On the other hand, Morioka *et al.* explored a colour-based model to identify object correspondences among different views (Morioka et al., 2006).

From previous works, it is clear that deep learning dramatically improves the performance for object detection. However, most of the existing object detection techniques focus on single-view images which can be affected by pose, occlusion, etc. In contrast, the multi-view object detection approach considered in this paper can address this problem by exploiting additional cues from other views. Further, multi-view tracking methods mostly rely on initial detection result and are limited to videos, making those techniques unsuitable for tasks lacking temporal information. We instead propose a method which can be used across views without temporal information. Our method leverages the superior detection performance achieved by recent single-view deep learning architectures and extends them to the multi-view domain through embedded components sharing information between pairs of views.

3 METHODOLOGY

In this paper, we propose to extend a pre-trained single-view detection network to a multi-view framework with embedded components for object detection and labelling, thus capitalising on the advantages of recent deep learning architectures while at the same time overcoming the shortage of labelled multi-view datasets that prevent direct training of multi-view architectures. Our framework is built based on Mask-RCNN (He et al., 2017) for extracting candidate bounding boxes. We enforce the epipolar geometry to constrain the locations of instances matched across two views, and compute an efficient person re-identification feature to measure the appearance similarity between matched instances. The Hungarian algorithm is used in combination with a confidence strategy to identify matching pairs and assign labels in polynomial time without the requirement to know the number of matching pairs. In our approach, we jointly optimise instance detection and labelling for optimal performance. In this section, we provide a description of the framework in the case of pairwise detection and labelling, leaving the generalisation to three or more views to future work.

3.1 Framework Overview

The proposed multi-view framework for object detection and labelling extends the classical Mask R-CNN approach by scaling it to multiple branches responsible for the processing of the different input views (two branches corresponding to two calibrated input views considered in this paper). The weights used for each branch of this extended framework are obtained from the original Mask R-CNN which was trained on the COCO dataset (Lin et al., 2014). The input to the network consists of two calibrated images of the scene captured from different viewpoints. The output for each input image consists of class, bounding box, mask and label information.

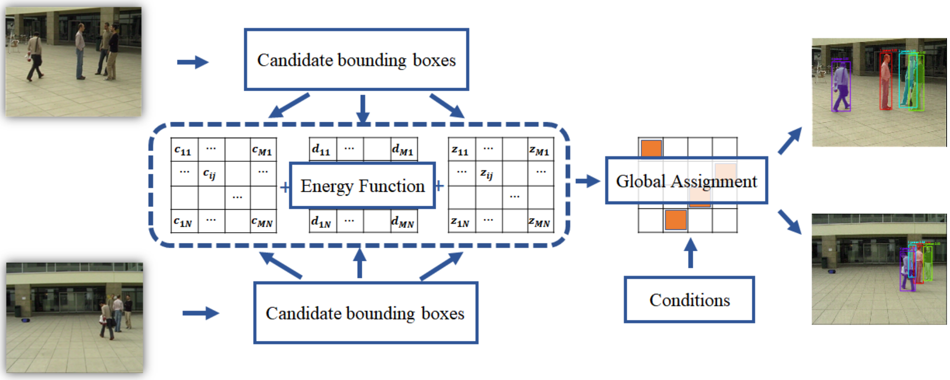

Figure 2: Overview of our proposed framework. First, candidate bounding boxes for each branch are extracted from a pair of images respectively utilising a pre-trained Mask R-CNN. Then based on these candidate bounding boxes, an energy function is defined by combining class, distance and appearance similarity matrices. Then the result characterising the class, bounding box, mask and label for pair of images are extracted by solving a constrained global assignment problem.

The two input images are first each fed to the deep learning network of their corresponding branch to extract convolutional features. The backbone for feature extraction is composed of a 50-depth Resnet and a feature pyramid network (Lin et al., 2017). Then a region proposal network generates candidate bounding boxes based on the convolutional features. A confidence score is obtained for each candidate bounding box. Then based on each candidate bounding box, the class is obtained using a fully-connected layer and the mask is obtained using a fully convolutional network (Long et al., 2015).

The key contribution is the introduction of embedded components which are used to link the two branches and enforce multi-view constraints. This is a critical step to ensure that the branches are able to cooperate when detecting and labelling objects. This is achieved by defining an energy function enforcing epipolar constraints, appearance similarity and class consistency amongst pairs of candidate bounding boxes between two input images. To efficiently solve the problem, this multi-view detection and labelling task is treated as a global optimisation problem with unknown number of assignments. The entire framework is illustrated in Figure 2.

3.2 Energy Formulation

Let us denote by I_1 and I_2 the two input images from two views fed to the pipeline. For each image, convolutional features are extracted by a 50-depth Resnet. Subsequently, based on the convolutional features, the region proposal network generates a number of candidate bounding boxes and corresponding confidence scores for each image. Each bounding box represents a candidate instance. Within each branch, these initial processing steps are similar to those in Mask-RCNN. Let us denote by M and N the number of candidate bounding boxes for the two images I_1 and I_2 respectively. The sets for candidate bounding boxes extracted from I_1 and I_2 are $B_1 = \{b_{11}, b_{12}, ..., b_{1i}, ..., b_{1M}\}$ and $B_2 = \{b_{21}, b_{22}, ..., b_{2j}, ..., b_{2N}\}$ respectively, where b_{1i} denotes the i-th instance in I_1 and b_{2j} denotes the j-th instance in I_2. To optimally match instances across the two views, we combine class, distance and appearance features to represent each instance.

Distance matrix: To measure the distance between two instances across the two views, first we represent the location of each instance with two points: the mid-point of the top line segment and the mid-point of bottom line segment of the corresponding bounding box. The location of an instance in I_1 can be represented as (x^t_{1i}, y^t_{1i}), (x^b_{1i}, y^b_{1i}) and similarly as (x^t_{2j}, y^t_{2j}), (x^b_{1j}, y^b_{1j}) in I_2. Then we use the epipolar geometry to measure the consistency between the two instances from I_1 and I_2. For a given point x in one view, there is an epipolar line $l = Fx$ in the other view on which the corresponding point x' must lie, where F is the fundamental matrix relating the two views (Hartley and Zisserman, 2003). F can be obtained either directly from the calibration information, if available, or otherwise indirectly by establishing a sufficient number of correspondences across views.

More specifically, for the i-th instance in I_1, the epipolar lines l^t_{1i}, l^b_{1j} in I_2 can be calculated as:

$$l^t_{1i} = F_{12}(x^t_{1i}, y^t_{1i}, 1)^\top, \quad (1)$$

$$l^b_{1i} = F_{12}(x^b_{1i}, y^b_{1i}, 1)^\top, \quad (2)$$

where F_{12} represents the fundamental matrix from I_1 to I_2. We denote by $D_{12}(i, j)$ the distance between

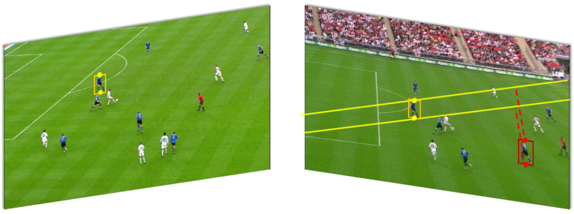

Figure 3: Epipolar geometry for distance measurement. The left image shows a bounding box in View 1 with the two points representing its location. The right image shows the two corresponding epipolar lines in View 2 inferred from the two points in View 1. In the right image, the instance in the yellow box denotes the correct correspondence while the instance in the red box corresponds to an incorrect match.

the top and bottom epipolar lines inferred from the i-th instance in I_1 and j-th instance in I_2. Conversely, $D_{21}(j,i)$ denotes the distance between the top and bottom epipolar lines inferred from the j-th instance in I_2 and i-th instance in I_1.

Then for each pair (i,j), the distance between the i-th instance in I_1 and the j-th instance in I_2 can be calculated as:

$$D(i,j) = D_{12}(i,j) + D_{21}(j,i), \tag{3}$$

where D denotes the distance matrix. An example for distance measurement can be seen in Figure 3. As illustrated in the figure, an object far from the corresponding epipolar lines will have a large distance value compared to the correct matching object.

Appearance similarity matrix: To match instances between two views, distance is not sufficient. An efficient feature that can properly represent appearance in multiple views is therefore necessary. For multi-view instance labelling, the viewpoint change can be quite large. The appearance feature should therefore be robust to large viewpoint change and variation in illumination. In our method, we use a feature called Local Maximal Occurrence (LOMO) feature and its metric learning for feature similarity measurement. It was first proposed by Liao *et al.* for person re-identification (Liao et al., 2015). The LOMO feature is an efficient feature combining a colour descriptor and a scale invariant local ternary pattern (SILTP). It uses a Retinex algorithm (Jobson et al., 1997) to handle illumination variation and extracts the maximal occurrence horizontally to overcome viewpoint change. In our method, we resize all the candidate bounding boxes to 128 × 48 as the same size used in person re-identification. Then the extracted LOMO feature for instances in I_1 can be represented as $H_1 = (h_{11}, h_{12}, ..., h_{1i}, ..., h_{1M})$ and $H_2 = (h_{21}, h_{22}, ..., h_{2j}, ..., h_{2N})$ in I_2. According to the metric learning in (Liao et al., 2015), the appearance similarity between two instances in two views can be calculated as:

$$Z(i,j) = (h_{1i} - h_{2j})^T W (\Sigma_I'^{-1} - \Sigma_E'^{-1}) W^T (h_{1i} - h_{2j}), \tag{4}$$

where Z denotes the appearance matrix, W represents a subspace, $\Sigma_I'^{-1}$ and $\Sigma_E'^{-1}$ represent the covariance matrices of intrapersonal variations and the extrapersonal variations respectively. The above-mentioned parameters used for appearance similarity calculation were set to the values originally proposed in (Liao et al., 2015), that is no additional training was performed. The value in Z will be low when two instances are similar.

Class similarity matrix: An additional consideration for multi-view tracking is that the two instances in I_1 and I_2 belonging to a pair should belong to the same class. The class for a candidate bounding box in I_1 can be represented as $C_1 = \{c_{11}, c_{12}, ..., c_{1i}, ..., c_{1M}\}$ and $C_2 = \{c_{21}, c_{22}, ..., c_{2j}, ..., c_{2N}\}$ in I_2. Therefore, we define the class matrix C as

$$C(i,j) = \begin{cases} 0 & \text{if } c_{1i} = c_{2j} \\ \sigma_{cls} & \text{if } c_{1i} \neq c_{2j}, \end{cases} \tag{5}$$

where σ_{cls} is used to penalise inconsistent class assignments.

Energy function: Finally we combine class, distance and appearance matrices as the representation of similarity between two instances in two views. Then the energy function for multi-view detection and tracking can be defined as E and be represented as:

$$E = w_\mathrm{D} \cdot D + w_\mathrm{Z} \cdot Z + w_\mathrm{C} \cdot C, \tag{6}$$

where w_D, w_Z and w_C are the weights for distance, appearance and class respectively. The weights w_D, w_Z and w_C are all set to 1 in this paper.

3.3 Global Optimisation

Matching pairs across the two views are extracted by minimising the energy function defined in the previous section. We regard this as an assignment problem as an instance in one view can only have at most one matched instance in the other view. However, this is not a classical assignment problem in the sense that the number of matching pairs are unknown. Therefore, to further constrain the problem, eligible pairs are required to satisfy the following two conditions.

Condition 1 (similarity): In practice, not all instances in one view will have a corresponding instance in the other view due to the visibility from different viewpoints. Therefore, a similarity threshold denoted by λ is introduced to prevent instances that are not sufficiently similar from being matched. This can remove candidate object pairs with large distance or appearance similarity. The value λ varies based on various factors such as instance resolution, illumination variation and distance between viewpoints. This condition is set to prescribe a minimum similarity to avoid selecting non-matching pairs. Therefore, we only consider the elements in the energy matrix E with values smaller than λ so that elements with values larger than λ are ignored.

Condition 2 (confidence score): In Mask R-CNN, each candidate bounding box is generated with a confidence score. We define the confidence score set as $S_1 = \{s_{11}, s_{12}, ..., s_{1i}, ..., s_{1M}\}$ corresponding to each bounding box in I_1 and $S_2 = \{s_{21}, s_{22}, ..., s_{2j}, ..., s_{2N}\}$ corresponding to each bounding box in I_2. Due to different resolutions and viewpoints, the same instance may have a low confidence score in one view while having a high confidence score in another view. Compared with detecting instances from a single view, multiple views can provide a richer source of information for detection. In particular, we consider pairs of scores jointly instead of single confidence scores. We define two confidence thresholds for each pair: β_n and β_h. β_n is identical to the confidence threshold used in Mask R-CNN and accounts for the possibility that an object may only be visible in one view. In contrast, the threshold β_h is introduced to help recover an object which may have a low confidence score in one view but a high confidence score in the other view. For each pair, we adopt the strategy that if the sum of confidence scores is higher than $2\beta_n$ or if either of the scores is larger than β_h, then the pair is regarded as eligible.

The global optimization problem can be formulated as an assignment problem with two additional conditions that are introduced to handle the unknown number of assignments. More specifically, this is achieved by finding the set of assignments $\mathbb{P}$ that minimises the previously defined energy function combined with a term aiming to maximise the number of assignments, namely

$$\begin{aligned} &\sum_{i,j \in \mathbb{P}} E(i,j) + (G - |\mathbb{P}|) \cdot \lambda \\ \text{s.t. } &(1)\ S_{1i} + S_{2j} > 2\beta_n \text{ or } S_{1i} > \beta_h \text{ or } S_{2j} > \beta_h \\ &(2)\ E(i,j) < \lambda \end{aligned} \tag{7}$$

where $G = \min\{M, N\}$ represents the maximum number of assignment in a $M \times N$ matrix, $|\mathbb{P}|$ denotes the cardinality of the set of chosen assignments (from a maximum of G possible assignments) and λ is the threshold on the pairwise similarity mentioned in condition 1. Solving this equation, we can correctly match objects in two views and extract the mis-detected bounding box in one view with the additional cues from the other view.

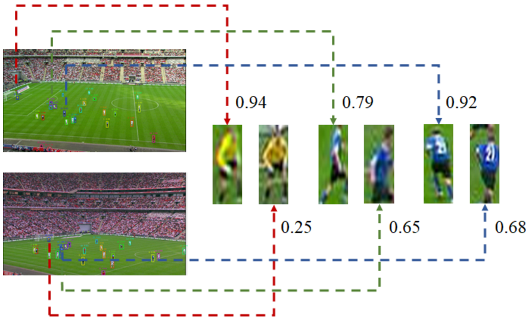

Figure 4: Illustration of the process to extract matching pairs in the presence of a low confidence score in one view in the case of the football dataset. All three example matching pairs shown are correctly detected across views despite the low confidence score in some of the views which would have resulted in mis-detections using the classical Mask R-CNN approach applied to each view individually.

We use the Hungarian algorithm to find the optimal assignment number $K = |\mathbb{P}|$ and the corresponding optimal assignments $\mathbb{P}$. The Hungarian algorithm is widely used in assignment problems as it achieves the optimal assignment with minimum cost. However, in this method, the number of potential pairs is unknown which means not all the assignments are eligible. This prevents direct application of the Hungarian algorithms which would also consider the pairs with values exceeding the threshold λ or having a low confidence scores when trying to achieve the minimum cost, thereby resulting in assignments which do not satisfy the imposed constraints.

To overcome this problem, we first cap all values in E at λ. Clamping the values ensure that the assignments corresponding to hypotheses that do not meet the constraints all bear the same penalty. Then we apply the Hungarian algorithm on the updated matrix and compute the optimal $G = \min\{M, N\}$ assignments by applying the Hungarian algorithm. From the G assignments, only assignments with a value smaller than λ and satisfying the confidence score mentioned in condition 2 qualify. Thus we remove the assignments which have a value λ or which do not satisfy the confidence score condition. We then obtain K assignments, each assignment representing a match between two views.

Finally, the single instances that only can be seen in one view with no corresponding pair in the other view are then extracted based on the confidence score β_n in the same manner as in the classical Mask R-CNN formulation. An example for pairs assignment with a low confidence score in one view is shown in Figure 4. It demonstrates that our system can correctly assign pairs even when one of them has a low confidence score.

Table 1: Details of the different datasets used for evaluation.

Datasets	View Number	Pairs of Images	Image Resolution
Campus	3	90	360 × 288
Terrace	4	180	360 × 288
Football	5	100	1920 × 1080
Basketball	4	180	360 × 288

4 EXPERIMENTAL EVALUATION

In this section, we start by introducing the datasets, the methods used for comparison and the metrics used for performance evaluation. Then we demonstrate the qualitative and quantitative performance for labelling multiple objects between views to show the advantages of the proposed multi-view framework in the context of a range of scenes with different degrees of complexity.

4.1 Evaluation Protocol

Datasets: Multi-view image datasets of scenes containing multiple objects are required for qualitative and quantitative evaluation. Due to the lack of dedicated multi-view image datasets, we instead use four challenging multi-view video datasets namely *Campus*, *Terrace*, *Basketball* and *Football* from which we extract a number of image pairs distributed across the sequences. The *Campus*, *Terrace* and *Basketball* datasets are from EPFL (Fleuret et al., 2008). For the *Campus* and *Terrace*, we use the corresponding first sequences. The *Football* dataset is from (Guillemaut and Hilton, 2011). These four datasets are challenging due to the wide baseline, severe occlusions, varying illumination and small object scale in some cases. These datasets are well suited to evaluated the performance and robustness of the proposed approach under operating conditions with varying degrees complexity.

The proposed method is applied on single frames. For a fair evaluation, each dataset is sampled using a fixed interval which is dependent on the length of the sequence. More specifically, we extract 10 frames from the shorter 100-frame *Football* video which contains five viewpoints. With $C_5^2 = 10$ possible camera pair combinations for each frame, this results in a total of 100 pairs of images in this dataset. Similarly, we extract 30 frames for each of the *Campus*, *Terrace* and *Basketball* resulting in a total of 90, 180 and 180 image pairs respectively. The details for these datasets are listed in Table 1. For each dataset, ground truth is generated by manually annotating the selected frames according to the guidelines from VOC2011 annotation (VOC,).

Comparison: Limited work has been conducted in multi-view detection and labelling in the image domain as most works tend to focus on video analysis and rely on the available temporal information. This severely limits the number of baseline approaches that can be used for comparative evaluation in the context of images only. Our approach is compared against two approaches: one based on a sequential tracking-by-detection approach; another one based on the probabilistic occupancy map approach (POM) (Fleuret et al., 2008). The tracking-by-detection approach uses the same components as in our proposed approach but in a sequential manner instead of the integrated framework we proposed. Instead of directly embedding components into the architecture to connect multi-view cues and jointly optimize detection and labelling between two views, the sequential approach first generates bounding boxes for each image in a pair using the classical Mask R-CNN approach before further processing to perform the labelling. We also compare our method with the well-established probabilistic occupancy map (POM) (Fleuret et al., 2008) approach which is applied to pairs of views using the default settings recommended by the authors. Object detection is restricted to the area of interest to allow a fair comparison with the POM method.

Evaluation metrics: Performance is evaluated based on the selected set of pairs of images for all datasets considered. In this paper, we use precision and recall to quantitatively evaluate the performance for multi-view multi-object labelling. Specifically, we first define the correct labelling number. To measure the correct labelling number, we calculate the intersection over union (IoU) between the detected bounding box and the ground truth for each instance. We denote by δ the threshold for the IoU with δ set to 0.5. Instances appearing in only one view are regarded as a correctly labelled object if the IoU is larger than δ. For objects appearing in both views, if the IoU values for both are larger than δ and they share the same label, we regard them both as correct. If they do not satisfy the above conditions or only satisfy one condition, we regard them both as false positive objects. Then, the precision is defined as the ratio of correctly labelled instances to all detected instances while the recall is defined as the ratio of correctly labelled instances to all ground truth instances. The quantitative result for each dataset is based on pairs of views in all datasets. We then use the mean value for all combinations of pairs which includes not only pairs defined by adjacent views but also between views in opposite directions.

Table 2: Quantitative result for multi-view tracking in the *Campus*, *Terrace*, *Basketball* and *Football* datasets.

	Campus		Terrace		Basketball		Football	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
POM (Fleuret et al., 2008)	72.01	70.26	67.81	49.90	59.04	22.21	49.93	17.51
Sequential framework	**95.65**	93.84	**79.26**	75.71	**77.42**	70.06	78.36	70.25
Proposed	95.25	**94.29**	79.24	**78.30**	73.26	**72.33**	**79.14**	**73.94**

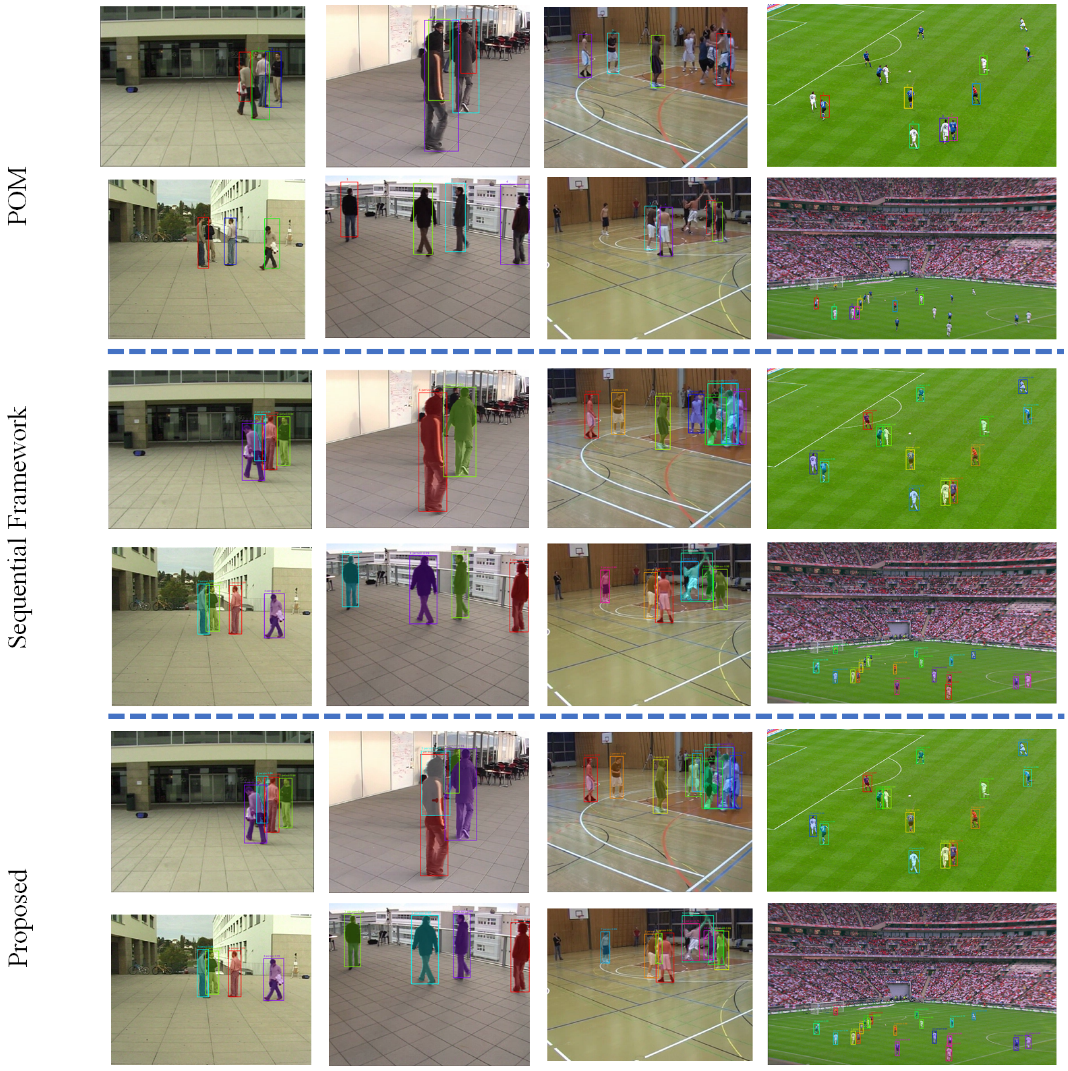

Figure 5: Qualitative result for the POM, the sequential approach and the proposed method. From left to right, we demonstrate the results in the *Campus*, *Terrace*, *Basketball* and *Football* datasets. For each pair of images, the instance with same label is represented in the same colour.

4.2 Quantitative Analysis

The quantitative results for multi-view labelling are listed in Table 2. In the implementation, we discard the candidate bounding boxes with confidence score lower than 0.1 before matching to improve the efficiency. The similarity threshold was set to $\lambda = 300$ for *Football*, $\lambda = 400$ for *Basketball* and $\lambda = 450$ for *Campus* and *Terrace*, these being mainly influenced by the baseline separating viewpoints and image resolution. From the table, we can see that the proposed method outperforms both the sequential and the POM approaches in all datasets in terms of the recall metric. This indicates that the proposed approach is able to recover objects which are otherwise undetected by the other approaches. In terms of precision, the proposed approach outperforms the POM approach on all datasets, while it performs overall similarly to the sequential approach. The gain in precision is most apparent in the case of the *Football* dataset, thus demonstrating the advantage of leveraging multi-view cues and closely coupling the detection and labelling as the complexity of the scene increases.

4.3 Qualitative Analysis

Qualitative results for multi-view labelling are shown in Figure 5. This provides an illustration of the performance of the proposed framework with outputs showing the class, bounding box, mask and label on the *Campus*, *Terrace*, *Basketball* and *Football* datasets, with two pairs of images provided for each dataset respectively. Results indicate that the proposed method performs well even under large viewpoint change, varying illumination and small instances. Moreover, the proposed method can also be applied in labelling objects in multiple classed by replacing person reidentification appearance feature with other features. We demonstrate some qualitative results on pairs of images including a variety of common object classes in Figure 6. The pair in first row in Figure 6 is extracted from the multi-view car dataset (Ozuysal et al., 2009), while the other three pairs of images were captured as part of this project.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we have extended a state-of-the-art single-view deep learning network into a joint multi-view detection and labelling framework without need for additional training. This is achieved by introducing a new architecture that extends a pre-trained network to multiple branches with additional components linking the different branches and enforcing multi-view constraints on the geometry, appearance and semantic content. By leveraging multi-view cues and closely integrating them into the proposed architecture, we demonstrate that it is possible to recover object instances which are otherwise hard to detected in single views. The proposed network has the added benefit of providing coherent multi-view labelling of the detected instances.

In future work, we will extend the method to multi-view videos. We anticipate that the temporal information present in multi-view videos will provide additional cues to resolve existing ambiguities and further improve object detection and labelling performance. Another interesting direction for future work would be to train an end-to-end deep neural network for multi-view object detection and labelling. This would remove the current reliance on heuristics, but would require access to large annotated multi-view datasets which is currently problematic.

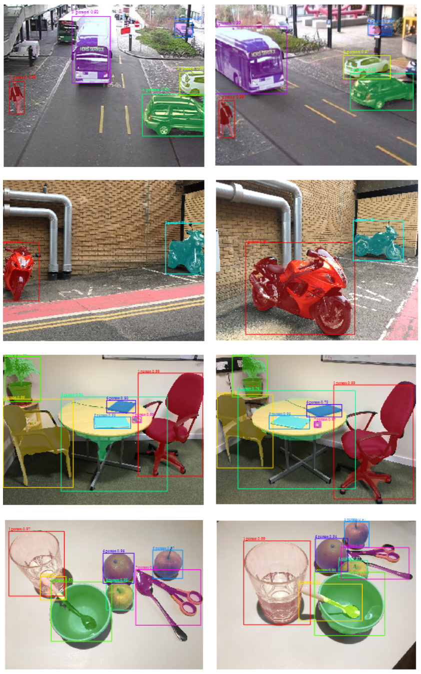

Figure 6: Results for the proposed method applied to pairs of images containing various common object classes.

ACKNOWLEDGEMENTS

This work was supported by the China Scholarship Council (CSC).

REFERENCES

VOC2011 annotation guidelines. `http://host.robots.ox.ac.uk/pascal/VOC/voc2011/guidelines.html`. Accessed: 06 Jun 2019.

Chang, T.-H. and Gong, S. (2001). Tracking multiple people with a multi-camera system. In *Proc. IEEE Workshop on Multi-Object Tracking*, pages 19–26. IEEE.

Dai, J., He, K., and Sun, J. (2015). Convolutional feature masking for joint object and stuff segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 3992–4000.

Erhan, D., Szegedy, C., Toshev, A., and Anguelov, D. (2014). Scalable object detection using deep neural networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 2147–2154.

Fleuret, F., Berclaz, J., Lengagne, R., and Fua, P. (2008). Multicamera people tracking with a probabilistic occupancy map. *IEEE trans. Pattern Analysis and Machine Intelligence*, 30(2):267–282.

Girshick, R. (2015). Fast R-CNN. In *Proc. IEEE International Conference on Computer Vision*, pages 1440–1448.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587.

Guillemaut, J.-Y. and Hilton, A. (2011). Joint multi-layer segmentation and reconstruction for free-viewpoint video applications. *International Journal of Computer Vision*, 93(1):73–100.

Hartley, R. and Zisserman, A. (2003). *Multiple view geometry in computer vision*. Cambridge university press.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask R-CNN. In *Proc. IEEE International Conference on Computer Vision*, pages 2961–2969.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE trans. Pattern Analysis and Machine Intelligence*, 37(9):1904–1916.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Hong, Z., Mei, X., Prokhorov, D., and Tao, D. (2013). Tracking via robust multi-task multi-view joint sparse representation. In *Proc. IEEE International Conference on Computer Vision*, pages 649–656.

Hosang, J., Benenson, R., Dollár, P., and Schiele, B. (2016). What makes for effective detection proposals? *IEEE trans. Pattern Analysis and Machine Intelligence*, 38(4):814–830.

Jobson, D. J., Rahman, Z.-u., and Woodell, G. A. (1997). A multiscale retinex for bridging the gap between color images and the human observation of scenes. *IEEE Trans. on Image processing*, 6(7):965–976.

Levine, S., Pastor, P., Krizhevsky, A., Ibarz, J., and Quillen, D. (2018). Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International Journal of Robotics Research*, 37(4-5):421–436.

Li, Y., Wang, S., Tian, Q., and Ding, X. (2015). Feature representation for statistical-learning-based object detection: A review. *Pattern Recognition*, 48(11):3542–3559.

Liao, S., Hu, Y., Zhu, X., and Li, S. Z. (2015). Person re-identification by local maximal occurrence representation and metric learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 2197–2206.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017). Feature pyramid networks for object detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer.

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440.

López-Cifuentes, A., Escudero-Viñolo, M., Bescós, J., and Carballeira, P. (2018). Semantic driven multi-camera pedestrian detection. *arXiv preprint arXiv:1812.10779*.

Luo, W., Xing, J., Milan, A., Zhang, X., Liu, W., Zhao, X., and Kim, T.-K. (2014). Multiple object tracking: A literature review. *arXiv preprint arXiv:1409.7618*.

Morioka, K., Mao, X., and Hashimoto, H. (2006). Global color model based object matching in the multi-camera environment. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2644–2649. IEEE.

Neelima, C., Harsh, A., Aroma, M., and Dhruv, B. (2015). Object-proposal evaluation protocol is 'gameable'. *CoRR*, abs/1505.05836.

Ozuysal, M., Lepetit, V., and Fua, P. (2009). Pose estimation for category specific multiview object localization. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 778–785. IEEE.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99.

Ristani, E. and Tomasi, C. (2018). Features for multi-target multi-camera tracking and re-identification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 6036–6046.

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., and LeCun, Y. (2013). Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*.

Szegedy, C., Reed, S., Erhan, D., Anguelov, D., and Ioffe, S. (2014). Scalable, high-quality object detection. *arXiv preprint arXiv:1412.1441*.

Xiao, T., Li, S., Wang, B., Lin, L., and Wang, X. (2017). Joint detection and identification feature learning for person search. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 3415–3424.

Xu, Y., Liu, X., Liu, Y., and Zhu, S.-C. (2016). Multi-view people tracking via hierarchical trajectory composition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 4256–4265.

Zhao, Z., Zheng, P., Xu, S., and Wu, X. (2018). Object detection with deep learning: A review. *CoRR*, abs/1807.05511.