A NOVEL MULTI-VIEW LABELLING NETWORK BASED ON PAIRWISE LEARNING

Yue Zhang, Akin Caliskan, Adrian Hilton, Jean-Yves Guillemaut

Centre for Vision, Speech and Signal Processing, University of Surrey, United Kingdom

ABSTRACT

Correct labelling of multiple people from different viewpoints in complex scenes is a challenging task due to occlusions, visual ambiguities, as well as variations in appearance and illumination. In recent years, deep learning approaches have proved very successful at improving the performance of a wide range of recognition and labelling tasks such as person re-identification and video tracking. However, to date, applications to multi-view tasks have proved more challenging due to the lack of suitably labelled multi-view datasets, which are difficult to collect and annotate. The contributions of this paper are two-fold. First, a synthetic dataset is generated by combining 3D human models and panoramas along with human poses and appearance detail rendering to overcome the shortage of real dataset for multi-view labelling. Second, a novel framework named Multi-View Labelling network (MVL-net) is introduced to leverage the new dataset and unify the multi-view multiple people detection, segmentation and labelling tasks in complex scenes. To the best of our knowledge, this is the first work using deep learning to train a multi-view labelling network. Experiments conducted on both synthetic and real datasets demonstrate that the proposed method outperforms the existing state-of-the-art approaches.

Index Terms— Multi-view network, synthetic dataset, multi-view labelling, multiple people labelling

1. INTRODUCTION

Multi-view multiple people labelling is a challenging task aiming to detect and label multiple people from different viewpoints in complex scenes. Existing 3D computer vision approaches such as shape reconstruction and depth estimation have been predominantly aimed at modelling a scene containing a single person [1, 2, 3]. However, real-world applications typically contain many people appearing simultaneously in the scene, often with some complex interactions. There are three main difficulties to address in multiple people multi-view labelling: First, building a suitable multi-view dataset with multiple people that can be leveraged to train dedicated architectures; Second, correctly detecting bounding boxes for multiple people in complex scenes; Third, finding reliable correspondences for multiple people from multiple viewpoints under wide baseline.

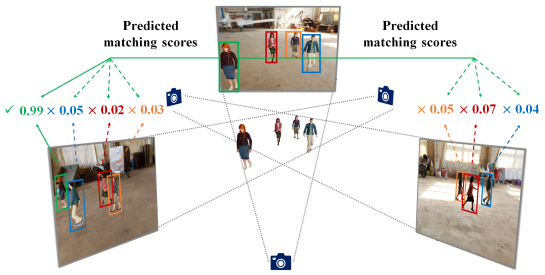

Fig. 1. Illustration of the proposed method. In the provided three views, the same person is labelled using a bounding box with consistent colour. A solid line indicates a correct matching pair (whose confidence score has been marked with a tick), while a dashed line indicates a non-matching pair (whose confidence score has been marked with a cross).

Although few approaches have directly tackled multi-view multiple people labelling, the related problem of identifying correspondences for people from multiple cameras has been explored in the person re-identification (re-id) task which is concerned with finding suitable features to represent appearance similarity across different cameras for multiple people. By using deep neural networks, the performance for person re-id has been improved in recent years [4, 5, 6, 7, 8]. However, these methods require correctly cropped bounding boxes with similar backgrounds seen from the different cameras. Besides, unlike the task considered in this paper, the bounding boxes from the images captured from different cameras do not relate to the same scene. More importantly, in person re-id, each person in the dataset bears a specific identity for the network to learn, while our task is concerned with consistently labelling people across views without prior information on their identities. Unlike person re-id which learns discriminative features with given IDs, the proposed method learns matching confidence only based on binary ground truth. Moreover, person re-id fails to exploit multi-view correspondences from cropped bounding boxes and non-overlapping cameras.

In multi-view tasks, conventional methods use multi-view geometry from calibration information such as ground plane homographies, epipolar lines or 3D positions to infer multi-view correspondences [9, 10, 11, 12]. However, when the detection is inaccurate or occlusions occur, geometry-based methods are prone to fail. Existing works on multi-view la-

Published in: 2021 IEEE International Conference on Image Processing (ICIP), pages 3682–3686, 2021. DOI: https://doi.org/10.1109/ICIP42928.2021.9506140

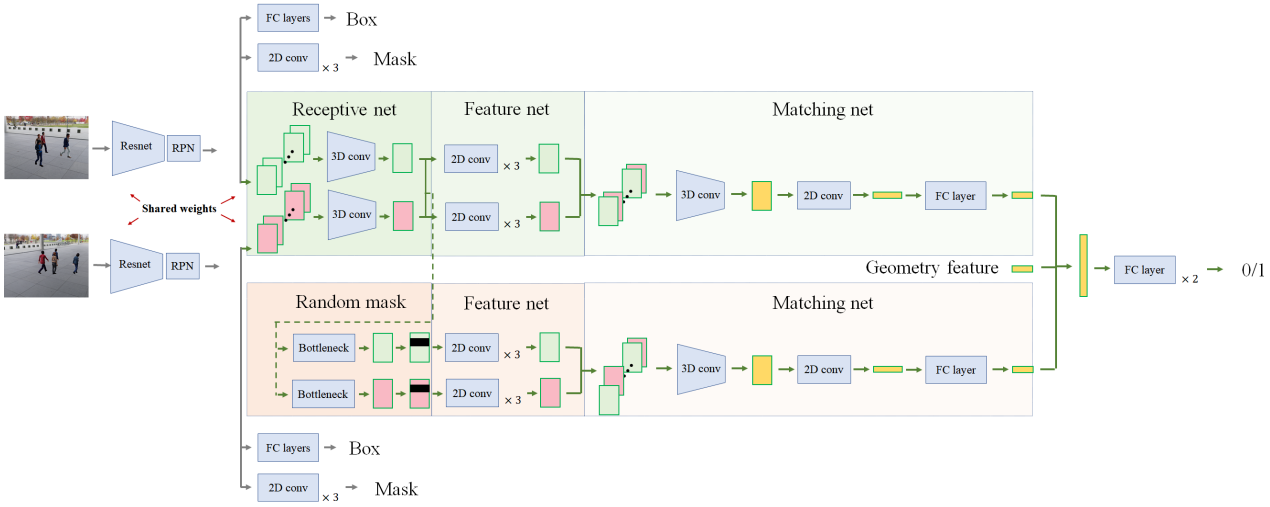

Fig. 2. Proposed MVL-net architecture. The proposed MVL-net consists of three main parts, namely, multi-view feature extraction (receptive net, feature net and random mask), calibration fusion (calibration vector and subsequent neural network), and matching net. The network also includes a global branch (shaded in green) and a local branch (shaded in red). The inputs for the network are pairs of images from two viewpoints, and the outputs are predicted matching confidence scores for instance pairs between two views.

belling and multi-view video tracking use either multi-view geometry constraints or pre-trained person re-id features to obtain the multi-view correspondences [13, 14, 15]. To the best of our knowledge, learning consistent feature representations for multi-view multiple people labelling from raw multi-view images has not been exploited in existing works. A core challenge which has prevented the training of multi-view networks is the lack of datasets for multi-view labelling. Unlike single-view data, multi-view data acquisition and ground truth labelling is a difficult and expensive process. Consequently, in the multi-view domain, synthetic data is widely used to overcome the shortage of real data and the associated difficulties.

In this paper, we propose a new approach to build a synthetic multi-view dataset and we introduce a novel architecture for multi-view multiple people labelling called MVL-net. The dataset is generated by combining 3D human models in various poses and background images along with a calibration setup and realistic rendering. Ground truth data including bounding box, segmentation, people correspondence and camera calibration are generated across all views. In the network, we combine several tasks including bounding box detection, semantic segmentation and matching prediction for multiple people in complex scenes. The contributions are as follows: First, we introduce the first large-scale synthetic multi-view dataset (*MV3DHumans*) for multi-view labelling of multiple people in various scenes. Second, a novel deep neural network that comprises a sophisticated design is proposed to unify the multi-view multiple people detection, segmentation and labelling tasks. Third, an evaluation is conducted to demonstrate the advance over the state of the art on both synthetic and real datasets.

2. METHODOLOGY

In this paper, we propose a unified framework for simultaneous multiple people detection, segmentation and labelling called MVL-net. The proposed MVL-net is based on image pairs acquired from different viewpoints, and can be trained using our proposed synthetic multi-view multiple people dataset to overcome the shortage of real-world training data.

2.1. Synthetic Multi-View Dataset Generation

To improve the generalisation of multi-view multiple people detection and labelling with respect to arbitrary human poses and actions, we introduced a new dataset, *Multi-View 3D Humans (MV3DHumans)*. This is the first large-scale multi-view multiple people dataset with synchronized cameras, suitable to train a detection and labelling network. Our approach achieves high-quality rendering by using Blender to combine synthetic human models with clothing and hair details, compositing them against different realistic backgrounds, and animating them to perform different actions.

The dataset is generated for various numbers of people (4, 6, 8 and 10) randomly positioned in the scene, using different 3D models of male and female characters with variations in clothing, pose and textures. In each frame, each scene is rendered into 16 camera views with in each case RGB images, instance segmentation masks, camera calibration and people correspondence. The *MV3DHumans* dataset is available at `https://cvssp.org/data/MV3DHumans`.

Fig. 3. Proposed *MV3DHumans* dataset generation framework. **a)** T-Pose human model generation, **b)** 3D models in action, **c)** Realistic rendering using environmental lighting, **d)** Images rendered from multiple viewpoints.

2.2. Multi-View Labelling Network (MVL-net)

The proposed MVL-net is built upon Mask R-CNN [16] generalised with a powerful labelling branch consisting of multi-view feature extraction, calibration fusion and matching net for multi-view multiple people matching. The number of people and their visibility from different viewpoints being unknown, it would be intractable to attempt to directly learn a fixed label for each person in the scene as done in person re-id. Instead, a matching confidence score is introduced to predict matching across multiple people. The architecture of the proposed multi-view network is shown in Figure 2.

Multi-View Feature Extraction To distinguish multiple people in the scene from different viewpoints, it is important to use a discriminative feature that is robust to viewpoint changes. Our network achieves this through the introduction of three key components: a receptive net, a random mask and a feature net. Specifically, to increase the receptive field of our MVL-net, we take advantage of multiple feature map outputs from the backbone and select the K region of interest (ROI) features with highest confidence scores. The top K feature maps, which capture a person's features in different shapes and details, are first fed into a 3D convolution layer to obtain a fine grained ROI feature. Then, 2D convolutions are performed to extract discriminative features for each person. Convolution weights are shared across views for ease of generalisation and computation. To further refine the features, a local branch is added to encourage the network to learn local and robust information, via use of a random mask as proposed in [8]. For multi-view feature learning, triplet loss is used as the loss function.

Calibration Component Calibration information can provide cues to infer correspondence between views. In the proposed method, a geometry feature vector obtained from calibration information is integrated into our MVL-net, so as to assist multi-view labelling. For each bounding box, points corresponding to its middle vertical are sampled to obtain a geometry distance representation for a pair. Moreover, the calibration information is also employed in the final prediction and decision for labelling each pair of people, based on the Euclidean distance between the estimated epipolar line and the corresponding point.

Matching Net Besides providing a measure of person similarity by extracting the feature distance, we utilise a matching net to accommodate features representing people from multiple views. For each pair of people, we first combine their features and obtain a 4-dimensional feature map, of size $C \times 2 \times W \times H$, with depth equal to 2, and C, W and H respectively denoting the number of channels, width and height. Assuming that there are N_1 people detected in View 1 and N_2 people in View 2, we obtain $N = N_1 \times N_2$ candidate pairs between the two views resulting in a combined feature map of size $N \times C \times 2 \times W \times H$. As an input to the matching net, the composed 3D convolution layer with a depth kernel of size 2 is used to obtain features for a pair. The weights are shared with the global and local branches. Then, global features, local features and geometry features from calibration component are aggregated as a pairwise similarity feature and then fed into two fully connected layers for matching confidence prediction. For matching confidence prediction, binary cross entropy is used as the loss function.

2.3. Multi-View Distance Measurement

To measure the multi-view distance for each pair, the epipolar geometry is also used to calculate the distance with regards to two key points, namely the top mid-point and the bottom mid-point. The distance similarity score for each point is defined as a piece-wise linear function which is, 1 when the distance is less than a threshold τ; 0 when the distance is larger than 5τ; and linear from 1 to 0 between τ and 5τ. The distance similarity score S_d for a pair is defined as an average of the distance similarity scores of the two key points. Denoting by S_m the predicted matching confidence score, the similarity score for label assignment to a pair is then obtained as the geometric mean $S_l = \sqrt{S_m \times S_d}$.

3. EXPERIMENTAL EVALUATION

3.1. Experimental Setting

Implementation Details The hyper-parameters and architecture for the backbone, bounding box head and mask head are identical to those used in Mask-RCNN-FCN. We use the model trained on the COCO person dataset as a pre-trained model. We train the MVL-net for $27,000$ iterations using synthetic image pairs, in which the batch size for both training and testing is set to 2.

Dataset In both training and testing datasets, 16 cameras are set up around the scene. The calibration parameters in training and testing datasets are identical. For training, 600 frames are generated based on 300 models and 35 panoramas. For testing, 300 frames are generated based on 100 models and 20 panoramas. The models and panoramas used for training and testing data generation are different.

Fig. 4. Qualitative results for multi-view labelling in 22.5° synthetic, 45° synthetic, 67.5° synthetic, 90° synthetic, campus, terrace and basketball datasets.

Baselines Due to the lack of approaches focusing on multi-view multiple people labelling, we compared our work with several recent deep learning based person re-id methods, namely, Batch DropBlock network (BDB) [8] and Relation-Aware Global Attention (RGA) [17]. As person re-id methods are based on cropped bounding box rather than raw images, we first extract bounding boxes using Mask R-CNN, and then predict the labelling through the person re-id feature distance. The Hungarian algorithm is then used to extract label assignments. We also compare our method with the traditional multi-view multiple people methods, Probabilistic Occupancy Map (POM) [18] and Trained Single-view Networks with Multi-view Constrains (SNMC) [15].

Metrics Multi-view labelling tasks focus on labelling people in images from two (or multiple) different viewpoints. Therefore, precision and recall are used to measure performance. To this end, we first define the correctly detected people in the scene given the ground truth and an overlap threshold of 0.5. Then, based on the correctly detected bounding box, the correct number of matching people is calculated. If a person is visible in one view but occluded (or out of camera range) in the other view, it is regarded as correctly labeled only if it has no matching person in the other view. Precision is then calculated as the ratio of the number of correctly labelled people N_c to the total number of detected people N_d, i.e. $Precision = N_c/N_d$, while recall is calculated as the ratio of the number of correctly labelled people N_c to the ground truth number of people N_{GT}, i.e. $Recall = N_c/N_{GT}$.

The Hungarian algorithm is used to identify correct assignments from all candidate pairs, with a minimum matching confidence score (set to 0.1) to filter invalid assignments due to some people being only visible in a sub-set of views.

3.2. Evaluation on Synthetic Data

Qualitative results are shown in the first two columns of Figure 4. From these examples, it can be observed that the proposed method is able to handle wide baselines and severe occlusions in crowded scenes. The quantitative results for synthetic data are reported in Table 1. These indicate that the proposed method outperforms all the baseline methods.

Table 1. Comparisons for multi-view multiple people labelling on the synthetic dataset (P=Precision, R=Recall).

Methods	22.5°		45°		67.5°		90°	
	P	R	P	R	P	R	P	R
POM [18]	46.86	26.58	32.10	19.17	26.07	16.16	24.49	15.68
SNMC [15]	83.18	86.73	81.16	84.89	80.18	83.97	79.72	83.49
BDB [8]	83.17	87.06	80.45	84.31	78.84	82.67	78.49	82.32
RGA [17]	83.25	87.14	80.55	84.43	79.14	82.97	78.64	82.47
Proposed	**94.31**	**95.17**	**93.73**	**94.68**	**93.85**	**94.98**	**93.75**	**94.75**

Table 2. Comparisons for multi-view multiple people labelling on the real datasets (P=Precision, R=Recall)

Methods	Campus		Terrace		Basketball	
	P	R	P	R	P	R
POM [18]	72.01	70.26	67.81	49.90	59.04	22.01
SNMC [15]	95.25	94.29	79.24	78.30	73.26	**72.33**
BDB [8]	92.05	92.09	70.78	69.67	59.09	55.37
RGA [17]	91.62	91.82	64.84	63.70	53.86	50.52
Proposed	**95.57**	**95.66**	**80.74**	**79.85**	**74.85**	69.54

3.3. Evaluation on Real Data

To validate our method on real data, we follow the work of SNMC [15] to uniformly sample 30 frames from campus (sequence1), terrace (sequence1) and basketball multi-view videos. The frames without people visible in all views are removed from the test. The quantitative results on real data are listed in Table 2. These indicate that the proposed trained network works well on challenging real data. Qualitative results are also shown in the last three columns of Figure 4.

4. CONCLUSIONS

In this paper, we have proposed a novel framework for dealing with multi-view multiple people labelling. This contributes a novel synthetic dataset with detailed human texture, pose and panoramic background, as well as, to the best of our knowledge, the first deep neural network for multi-view labelling. The effectiveness and efficiency of the proposed MVL-net have been validated on both synthetic and real datasets. Future work will investigate how keypoints and video information can be leveraged to further improve performance.

5. REFERENCES

[1] Marc Habermann, Weipeng Xu, Michael Zollhofer, Gerard Pons-Moll, and Christian Theobalt, "DeepCap: Monocular human performance capture using weak supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5052–5063.

[2] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll, "Video based reconstruction of 3D people models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8387–8397.

[3] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese, "3D-R2N2: A unified approach for single and multi-view 3d object reconstruction," in *European conference on computer vision*. Springer, 2016, pp. 628–644.

[4] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 274–282.

[5] Hong-Xing Yu, Wei-Shi Zheng, Ancong Wu, Xiaowei Guo, Shaogang Gong, and Jian-Huang Lai, "Unsupervised person re-identification by soft multilabel learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2148–2157.

[6] Yutian Lin, Lingxi Xie, Yu Wu, Chenggang Yan, and Qi Tian, "Unsupervised person re-identification via softened similarity learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3390–3399.

[7] Lingxiao He, Jian Liang, Haiqing Li, and Zhenan Sun, "Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7073–7082.

[8] Zuozhuo Dai, Mingqiang Chen, Xiaodong Gu, Siyu Zhu, and Ping Tan, "Batch DropBlock network for person re-identification and beyond," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 3691–3701.

[9] Mustafa Ayazoglu, Binlong Li, Caglayan Dicle, Mario Sznaier, and Octavia I Camps, "Dynamic subspace-based coordinated multicamera tracking," in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 2462–2469.

[10] Saad M Khan and Mubarak Shah, "A multiview approach to tracking people in crowded scenes using a planar homography constraint," in *European Conference on Computer Vision*. Springer, 2006, pp. 133–146.

[11] Laura Leal-Taixé, Gerard Pons-Moll, and Bodo Rosenhahn, "Branch-and-price global optimization for multi-view multi-object tracking," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[12] Kyungnam Kim and Larry S Davis, "Multi-camera tracking and segmentation of occluded people on ground plane using search-guided particle filtering," in *European Conference on Computer Vision*. Springer, 2006, pp. 98–109.

[13] Francois Fleuret, Jerome Berclaz, Richard Lengagne, and Pascal Fua, "Multicamera people tracking with a probabilistic occupancy map," *IEEE trans. Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 267–282, 2008.

[14] Yuanlu Xu, Xiaobai Liu, Yang Liu, and Song-Chun Zhu, "Multi-view people tracking via hierarchical trajectory composition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4256–4265.

[15] Yue Zhang, Adrian Hilton, and Jean-Yves Guillemaut, "A new approach combining trained single-view networks with multi-view constraints for robust multi-view object detection and labelling," in *VISIGRAPP*, 2020, pp. 452–461.

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, "Mask R-CNN," in *Proc. IEEE International Conference on Computer Vision*, 2017, pp. 2961–2969.

[17] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen, "Relation-aware global attention for person re-identification," *CVPR*, 2020.

[18] Francois Fleuret, Jerome Berclaz, Richard Lengagne, and Pascal Fua, "Multicamera people tracking with a probabilistic occupancy map," *IEEE transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 267–282, 2007.