Image and Vision Computing 88 (2019) 120–131

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Spatio-temporal object detection by deep learning: Video-interlacing to improve multi-object tracking☆

Ala Mhalla[a,*], Thierry Chateau[a], Najoua Essoukri Ben Amara[b]

[a]Université Clermont Auvergne, Institut Pascal, F-63000 Clermont-Ferrand, France
[b]LATIS ENISo, National Engineering School of Sousse, University of Sousse, Tunisia

ARTICLE INFO

Article history:
Received 1 March 2019
Accepted 5 March 2019
Available online 28 March 2019

Keywords:
Multi-object tracking
Interlacing and inverse interlacing models
Specialization
Interlaced deep detector

ABSTRACT

Tracking-by-detection have become a hot topic of great interest to some computer vision applications in the recent years. Generally, the existing tracking-by-detection frameworks have difficulties with congestion, occlusion, and inaccurate detection in crowded scenes. In this paper, we propose a new framework for Multi-Object Tracking-by-Detection (MOT-bD) based on a spatio-temporal interlaced encoding video model and a specialized Deep Convolutional Neural Network (DCNN) detector. The spatio-temporal variation of objects between images are encoded into "interlaced images". A specialized "interlaced object" convolutional deep detector is trained to detect objects in interlaced images and a classical association algorithm to perform the association between detected objects, since interlaced objects are built to increase overlap during the association step which leads to improve the MOT performance over the same detector/association algorithm applied on non-interlaced images.

The effectiveness and robustness of this contribution is demonstrated by experiments on popular tracking-by-detection datasets and benchmarks such as the PETS, TUD and the MOT17 benchmark. Experimental results demonstrate that interlacing video idea has many advantages to improve the tracking performances in terms of both precision and accuracy of tracking and illustrate that the "power of video-interlacing" outperforms several state-of-the-art tracking frameworks in multiple object tracking.

1. Introduction

Visual tracking for multiple objects in video sequences aims to estimate the trajectories of moving objects in a video sequence, and it has been extensively studied for several practical applications such as road traffic control, driving assistance, behavior analysis and video surveillance [1,2].

Generally, multi-target tracking is formulated as a data association task where a generic detector localizes object bounding boxes in each frame and then a data association algorithm associates corresponding detection boxes across frames for tracking purposes. Tracking approaches can be performed offline [3,4] by simultaneously exploiting all the object bounding boxes of processed images, or online [5,6] by limiting themselves to past images. The online approaches are selected when the real-time aspect is paramount and produce results that are fairly comparable to offline approaches as detailed in some studies [7,8]. Thanks to their superior performance and computation efficiency in object tracking, we are interested in the online tracking approaches and particularly on the problem of Multi-Object Tracking-by-Detection (MOT-bD).

However, the performance of the existing MOT-bD methods depends much on the quality of the initial detection and it is notable that a MOT-bD method becomes more reliable if it has a robust detector.

In recent years, deep learning techniques have achieved the state-of-the-art performance in several computer vision applications such as object detection [9,2], semantic segmentation [10,11] and object tracking [12,13]. Since recent object detectors based on Deep Convolutional Neural Networks (DCNN) have high performances, recent tracking approaches have mostly followed MOT-bD techniques: it consists in using a DCNN detector of the tracked-object classes to estimate the positions of the targets at each frame, following by an association step. Most popular MOT-bD methods include Faster R-CNN [14], SSD [15], or YOLOv3 [16] due to their performance and efficiency in detecting general objects.

Several approaches based on MOT-bD theories have been proposed by research groups in the world to solve the problems of multi-object tracking [17,18].

Among these approaches, we quote in particular the tracklet algorithms [19,17], which take a sequence of frames with their respective detections. Afterward, a tracklet association method associates

☆ This paper has been recommended for acceptance by Sinisa Todorovic.

* Corresponding author.
E-mail addresses: ala.mhalla@uca.fr (A. Mhalla), thierry.chateau@uca.fr (T. Chateau), najoua.benamara@eniso.rnu.tn (N. Essoukri Ben Amara).

https://doi.org/10.1016/j.imavis.2019.03.002
0262-8856/

target objects in a video sequence. Other tracking algorithms have been focused on using appearance models to estimate the tracks in each frame. In most cases, these models are learned online and utilized to estimate an affinity score for any track-detection pair. Some of the most popular work has suggested using more reliable and robust appearance models in order to differentiate objects of similar appearance [20,7].

Yet, the performances of the existing MOT-bD frameworks [19,21] are limited and still suffer from various challenges such as noisy detection, occlusions, inter and intra-class variation and inaccurate detection in crowded scenes, e.g., in many real scenarios when people walk side-by-side across a pedestrian crossing, a significant number of person will be occluded by 50% and more for the entire sequence. Such issues frequently affect the MOT-bD performance in real-world scenarios.

To address these problems, this paper proposes to build an interlaced representation of an input video sequence that combines several frames in an interlaced one. The resulting interlaced representation provides for each frame spatio-temporal information that should be learnt into a DCNN. By detecting moving objects, the network learns to associate several temporal instances of the object in the same interlaced frame. The produced detection is then easier to associate because it is naturally overlapped between successive interlaced frames. A global synoptic of the proposed MOT-bD framework is illustrated in Fig. 1.

Since interlaced objects are built to increase overlap during the association step which leads to improve the MOT performance over the same detector/association algorithm applied on non-interlaced images. The suggested framework proposes some advantages and several improvements to resolve tracking challenges, among the advantage of interlacing, for example, reducing the occlusion problem. Fig. 2 shows the efficiency of the suggested tracking framework to solve occlusion and intersection situations.

The main contributions of this paper can be summarized as:

- An original interlaced model combined with a specialized deep detector which improves the performances of multi-object tracking.
- A set of comparative experiments on the PET2009, TUD and MOT 2017 benchmarks, which achieves competitive results with current state-of-the-art tracking frameworks.

The structure of the paper is organized as follows. Section 2 provides the related work performed in the field of object tracking. A detailed description of our tracking framework is presented in Section 3. Section 4 describes the experimentation details and provides the experimental results. Finally, a conclusion is given in Section 5.

2. Related work and literature analysis

In this section, we are interested in the related multi-object tracking frameworks proposed to automatically track objects in video sequences.

In the recent years, multi-object tracking has attracted a lot of research groups in developing state-of-the-art theories and novel applications in several domains like robotics, video surveillance and intelligent transportation systems [22,12,21]. Several MOT-bD frameworks have been suggested by research community in the world to solve the problems of multi-object tracking [23–25,18]. Some of them have been based on recursive Bayesian filters such as the Kalman filter [26] and the Sequential Monte Carlo one [27] in order to handle data association problems.

Other recent approaches have been based on matching object hypotheses obtained by detection between two consecutive frames using their characteristics like the size, the representation, the appearance and the position [28,29,13]. On the other hand, tracking

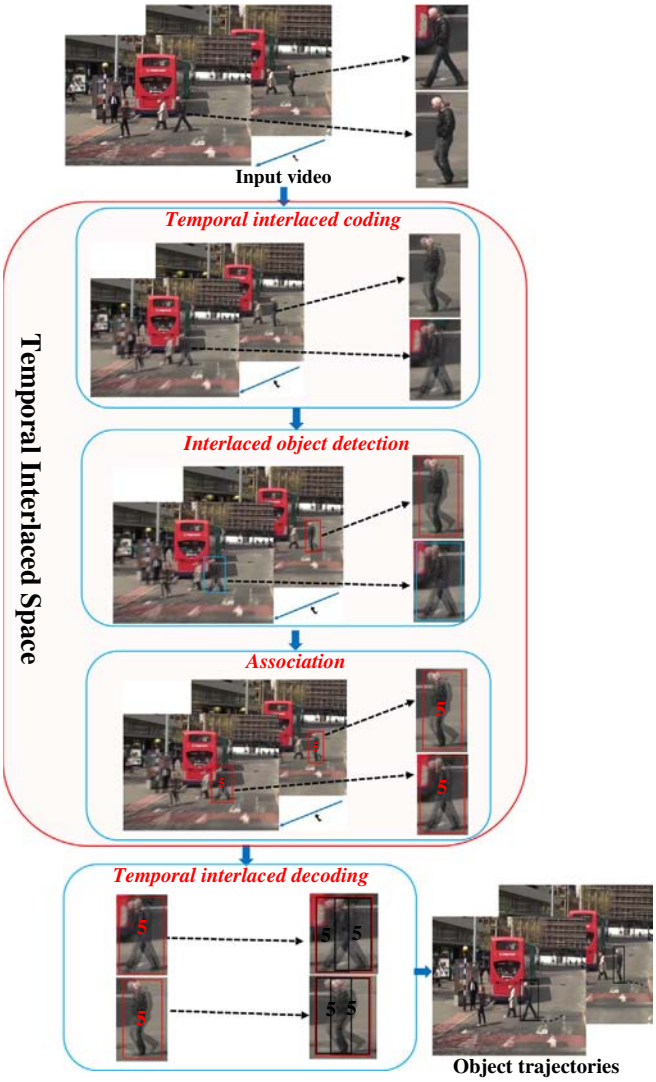

Fig. 1. General synoptic of proposed framework. Given a video taken by a mobile or stationary cameras, a temporal interlaced coding video model is applied to produce a set of interlaced images. Then a specialized DCNN detector is used to produce a list of detection from the interlaced images. Next, the detected objects are associated with a classical data association algorithm. Finally, a spatio-temporal interlaced decoding model is utilized to extract final trajectories into the initial video-sequence.

frameworks based on local data association (between two consecutive frames) have had critical limitations in resolving occlusion problems and therefore tend to generate short trajectories. Differently, some multi-object tracking frameworks build a set of trajectories through global or delayed optimization [7,30] in order to handle occlusion problems and noisy detection in tracking sequences.

Cox and Hingorani [31] suggested a classic Multiple Hypothesis Tracking (MHT), in an effort to delay association decisions until they were resolved. However, the number of hypotheses grew exponentially. An improved version of the MHT was proposed by Han et al. [32] which incorporated an appearance model to solve this issue. Kim et al. [28] introduced a revisited version of the standard MHT by including an online appearance model for multiple hypothesis tracking. This new formulation led to prove substantial performance gains over the old versions of the MHT by generating tracking hypotheses at each image with a prediction training model. The tracking hypotheses would conflict when attempting to assign different identifiers to the same object. The resolution of these conflicts generated global hypotheses with associated scores, and the best hypothesis was chosen to produce the final result.

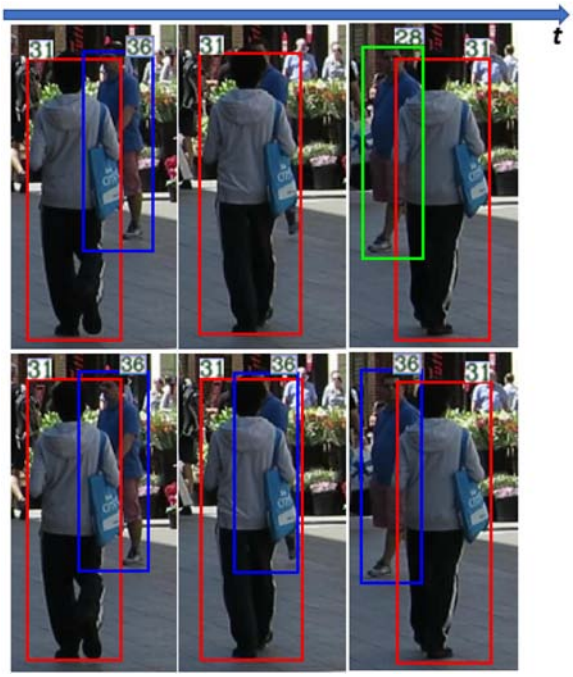

Fig. 2. Comparison of tracking with/without interlacing model in occlusion and intersection situations. The top row images show the result of the baseline framework (without interlacing) and the bottom ones present the result of the proposed one (with interlacing). The person (blue blobs) in the bottom row images is still detected and its identifier is preserved in occlusion situation. Contrarily, the person in the top row images is no longer detected and tracked (green blob). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Other methods used an appearance model or features for tracking. Danelljan et al. [33] put forward an on-line tracking framework based on adaptive color channels. This framework resolved several types of real-word scenarios, but it failed at scaling. Thus, the authors in [6] proposed solution to this problem.

Chari et al. [34] put forward pairwise costs to reinforce a min-cost network flow framework, which effectively handled overlapping problems and tracking enhancements. McLaughlin et al. [35] improved this min-cost network flow algorithm so that the tracking problem could be reduced in two steps. Firstly, an initial result was estimated without motion information, and secondly, it was then combined with a motion feature to generate a more reliable tracking solution.

Other state-of-the-art tracking frameworks have aimed to associate the detections by introducing a similarity function between detections based on the DCNN. The Triplet network mentioned in [36,37] and Siamese network [38] were efficiency techniques to measure the similarity between two objects. The Siamese network utilized a contrastive loss function to train the neural network, which helped the network to have small distances between the pair detections that belonged to the same objects while forcing the object with different identities to have large distances. This network is used to face recognition [39,40], single object tracking [41] and multi-target tracking [12,42]. The Triplet network, an enhanced version of Siamese network, was more robust to intra-class variations [37], since it used a new loss function for network training. It was utilized for characteristic learning 37,43] and object re-identification [44].

Accordingly, Wang et al. [19] proposed an original deep model for tracklet association that could join the Siamese CNN network learning and the temporal metrics so as to improve tracklet models.

Our approach is similar in spirit to the work of Tang et al. [45], which propose a new joint people detector that combines a state-of-the-art single person detector with a detector for pairs of people, which explicitly exploits common patterns of person–person occlusions across multiple viewpoints. In this direction, Tang et al. [45] propose a new double-person detector that allows to predict bounding boxes of two people even when they occlude each other by 50% or more, and propose a new training method for this detector.

Differently from the related work, we propose to enhance the performance of classical data association algorithms and help to recover objects in complex videos including occlusion, strong motion and intersection, by using an interlacing intermediate video representation model as well as a specialized DCNN detector.

3. Multi-object detection and tracking using interlaced video

This section presents the principle of the multi-object detection and tracking using interlaced videos.

A block diagram of the suggested framework is illustrated in Fig. 3. Given a video input taken by a stationary or mobile cameras, an interlacing model is applied to create an intermediate set of interlaced images. Then, a specialized DCNN detector fine-tuned by interlaced datasets provides objects on each interlaced frame. Targets (object trajectories) are produced by a classical association algorithm from detection. Finally, a reverse interlacing model is applied to extract final trajectories into the initial video sequence.

In what follows, we describe in details the main steps of the suggested tracking framework. In Subsection 3.1, we describe the proposed interlacing and inverse interlacing models. Subsection 3.2 shows how to build the interlaced DCNN detector.

3.1. Interlacing and inverse interlacing models

We present the interlacing and inverse interlacing mathematical models that serve as a basis of this work.

Given a set of temporal images $\mathcal{I} \doteq \{\mathbf{I}_k\}_{k=1,\dots K}$ extracted from the input video sequence, we propose to build an interlaced image set $\tilde{\mathcal{I}} \doteq \left\{\tilde{\mathbf{I}}_{\tilde{k}}\right\}_{\tilde{k}=1,\dots \tilde{K}}$. The tilde (~) notation is used for variables related to the interlaced video/image. If $\mathbf{I}_k(x,y)$ is the value of the (x,y) pixel (gray level or color) of an image k, the interlaced set of images are generated by Eq. (1):

$$\tilde{\mathbf{I}}_{\tilde{k}}(x,y) \doteq \sum_{d=0,\dots,(D-1)} \mathbf{I}_{(kg+ds)}(x,y).\delta(\,y[D]-d) \tag{1}$$

- $\delta(.)$ represents the Kronecker Delta function (2):

$$\delta(n) = \begin{cases} 1 \text{ for } & n=0 \\ 0 & \textit{otherwise} \end{cases} \tag{2}$$

- $y[D]$ is the modulo operator (y mod D), **D** is the depth (number of images used to build one interlaced image), **g** is a global step (difference between two successive interlaced images), and **s** is a local step (the gap between the frames which are combined for an interlaced image). Fig. 4 depicts the interlacing step.

Several configurations can be proposed. The key idea is that good configurations should produce a high overlap between detection. Fig. 5 shows examples of interlaced configurations for several sets of parameters (D,s,g).

Since the aim of this work is to detect and track objects, we define the bounding box associated to the object i at frame k by Eq. (3):

$$o_k^i \doteq \left(p_k^{(i,1)T}, p_k^{(i,2)T}, p_k^{(i,3)T}, p_k^{(i,4)T}\right)^T \tag{3}$$

where $p_k^{i,\cdot} \doteq \left(x_k^{(i,.)}, y_k^{(i,.)}\right)^T$ is the position of the four corners (upper left, upper right, lower right and lower left) of the bounding box.

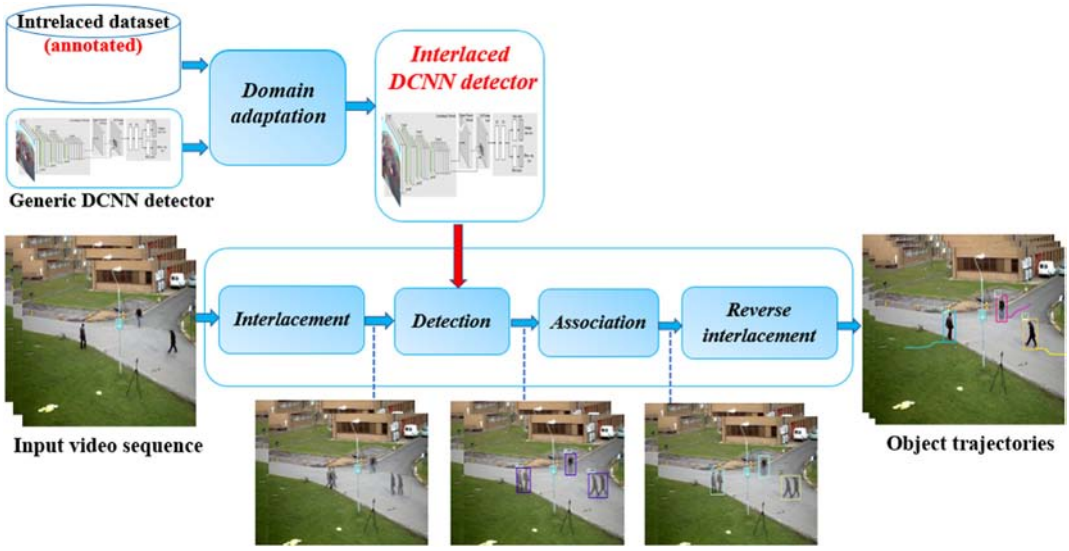

Fig. 3. Block diagram of the proposed framework: Given a video sequence, the suggested framework uses an interlacing step to create interlaced images. A generic deep detector adapted by interlaced datasets provides objects for each interlaced image. Next, a data association algorithm links detection on consecutive interlaced frames. Finally, estimated trajectories of objects are produced from interlaced ones by an inverse interlacing step.

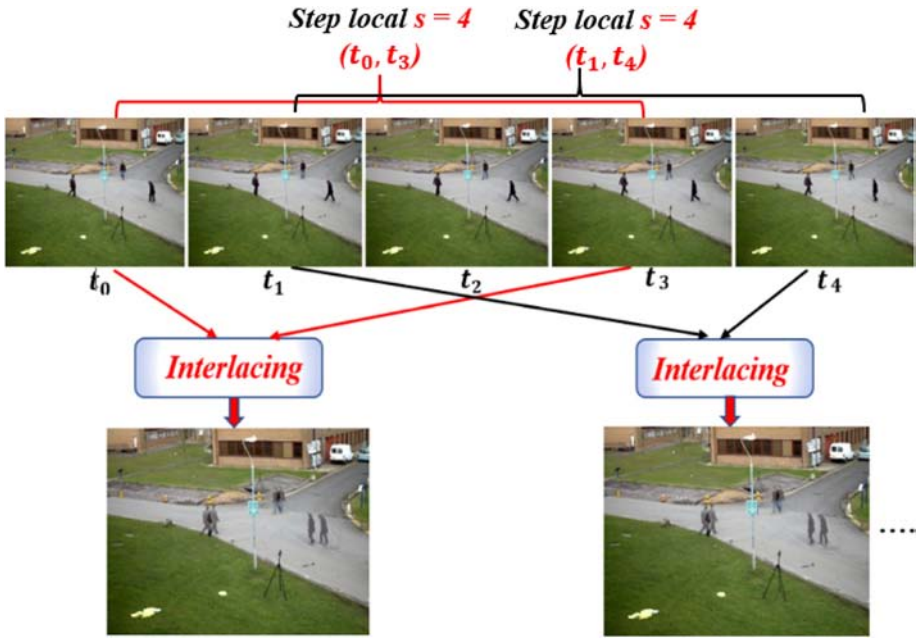

Fig. 4. Interlacing step. The top images are temporal frames extracted from a video sequence. The bottom images present results from interlaced model and combine several video frames: objects appears several times in interlaced image. In this example, the interlacing parameters (D, s, g) is fixed to $D = 2$ as the number of images in one interlaced image, $s = 4$ as the gap between the frames which are combined for an interlaced image, and $g = 1$ as the difference between two successive interlaced images.

Similarly, let us define the bounding boxes extracted from a tracking process applied on the interlaced video by Eq. (4):

$$\tilde{o}^i_{\tilde{k}} \doteq \left(\tilde{p}^{(i,1)T}_{\tilde{k}}, \tilde{p}^{(i,2)T}_{\tilde{k}}, \tilde{p}^{(i,3)T}_{\tilde{k}}, \tilde{p}^{(i,4)T}_{\tilde{k}}\right)^T \tag{4}$$

The object bounding box o^i_k is associated to the interlaced bounding box $\tilde{o}^i_{\tilde{k}}$ with Eq. (5):

$$\tilde{k} = \lfloor k/g \rfloor \tag{5}$$

where $\lfloor x \rfloor$ represents the floor function that takes as input a real number x and gives as output the greatest integer less than or equal to x.

Some interlaced configurations may produce interlaced images with a high redundancy; i.e., one original object for a given time can be extracted from several interlaced images. In this case, an average object position can be computed.

We consider a constant object velocity between two interlaced images. For a depth D, an interlaced image encodes objects D times. An estimation of the object bounding box o^i_k in the original image k can be extracted by interpolation between the interlaced bounding box $\tilde{o}^i_{\tilde{k}}$ and the interpolated interlaced bounding box $\tilde{o}^i_{\tilde{k}+\alpha_k}$. Fig. 6

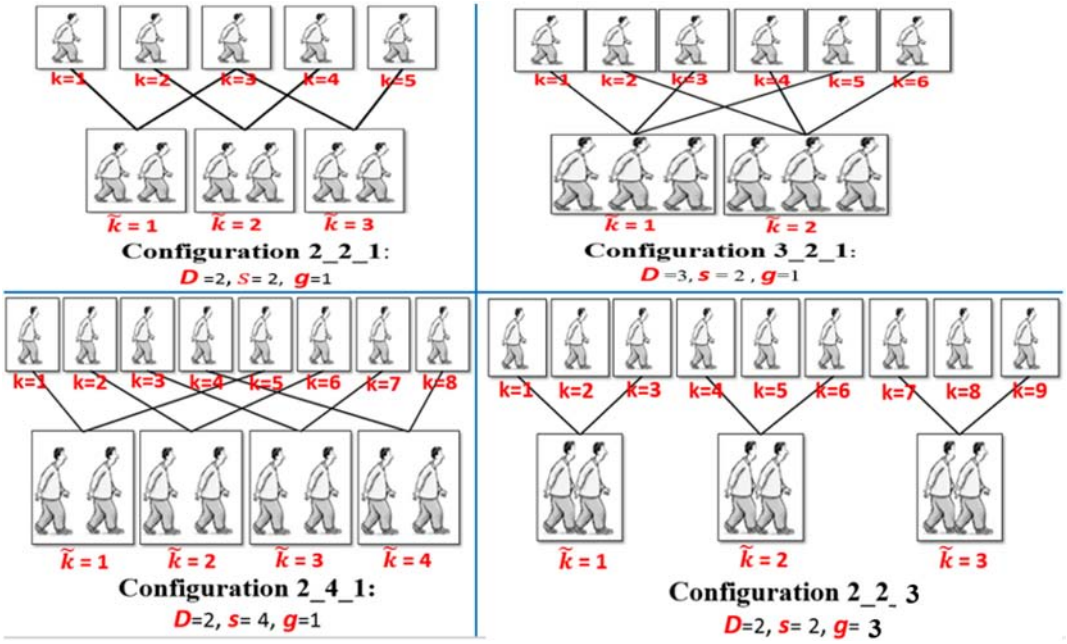

Fig. 5. Examples of interlaced configurations with four sets of parameters (D, s, g).

illustrates the estimation of $\tilde{o}^i_{\tilde{k}+\alpha_k}$. The latter box is computed by Eq. (6):

$$\tilde{o}^i_{\tilde{k}+\alpha} = \tilde{o}^i_{\tilde{k}} + \alpha\Delta_{\tilde{o}^i_{\tilde{k}}} \tag{6}$$

with:

$$\begin{cases} \alpha \doteq \frac{s(D-1)}{g} \\ \quad and \\ \Delta_{\tilde{o}^i_{\tilde{k}}} = \tilde{o}^i_{\tilde{k}+1} - \tilde{o}^i_{\tilde{k}} \end{cases} \tag{7}$$

where $\Delta_{\tilde{o}^i_{\tilde{k}}}$ is the displacement of interlaced bounding boxes between two successive interlaced frames.

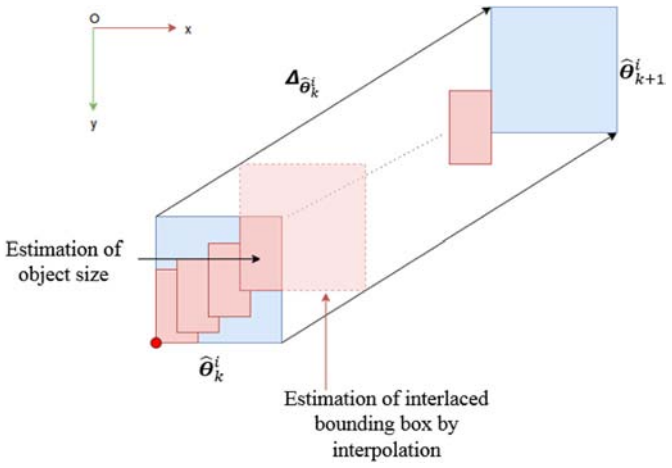

Fig. 6. Estimation of bounding boxes by interpolation model. The interpolation makes it possible to determine the size of an object that is assumed to be constant.

The next step consists in extracting the object bounding box $o^i_{g\tilde{k}+D}$ from the intersection between the interpolated object $\tilde{o}^i_{\tilde{k}+\alpha}$ and $\tilde{o}^i_{\tilde{k}}$ $(k = g\tilde{k})$, according to Eq. (8):

$$o^i_{g\tilde{k}+D} = \tilde{o}^i_{\tilde{k}} \cap \tilde{o}^i_{\tilde{k}+\alpha} \tag{8}$$

$\cap$ is the intersection operator between the two detections $o^i_{k_\cap} = o^i_{k_1} \cap o^i_{k_2}$ defined by Eq. (9):

$$\begin{cases} x^{(i,1)}_{k_\cap} = \max\left(x^{(i,1)}_{k_1}, x^{(i,1)}_{k_2}\right) \\ x^{(i,2)}_{k_\cap} = \min\left(x^{(i,2)}_{k_1}, x^{(i,2)}_{k_2}\right) \\ y^{(i,3)}_{k_\cap} = \max\left(y^{(i,3)}_{k_1}, y^{(i,3)}_{k_2}\right) \\ y^{(i,4)}_{k_\cap} = \min\left(y^{(i,4)}_{k_1}, y^{(i,4)}_{k_2}\right) \end{cases} \tag{9}$$

with $x^{(i,4)}_{k_\cap} = x^{(i,1)}_{k_\cap}, x^{(i,3)}_{k_\cap} = x^{(i,2)}_{k_\cap}, y^{(i,2)}_{k_\cap} = y^{(i,1)}_{k_\cap}$ and $y^{(i,4)}_{k_\cap} = y^{(i,3)}_{k_\cap}$.

In the same way, the object $o^i_{g\tilde{k}}$ is extracted from the intersection between $\tilde{o}^i_{\tilde{k}-\alpha}$ and $\tilde{o}^i_{\tilde{k}}$, according to Eq. (10):

$$o^i_{g\tilde{k}} = \tilde{o}^i_{\tilde{k}} \cap \tilde{o}^i_{\tilde{k}-\alpha} \tag{10}$$

Object ROIs for $k \in \left\{g\tilde{k}, g\tilde{k}+1, .., g\tilde{k}+D\right\}$ $\left(\tilde{k} = \lfloor k/g \rfloor\right)$ are estimated using linear interpolation Eq. (11):

$$o^i_k = o^i_{g\tilde{k}} + \beta\Delta_{o^i_{g\tilde{k}}} \tag{11}$$

with:

$$\begin{cases} \beta = k - g\tilde{k}Ds \\ \quad and \\ \Delta_{o^i_{g\tilde{k}}} = o^i_{g\tilde{k}+D} - o^i_{g\tilde{k}} \end{cases} \tag{12}$$

Fig. 7. Building an annotated interlaced video: Two images extracted from an interlaced video ($D = 2$) with object bounding boxes in black and interlaced bounding boxes in color. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.2. Interlaced deep detector

- **Interlaced dataset:** Given an annotated video, in which the trajectory of each object is labelized by a set of bounding boxes, we generate a new interlaced video dataset $\widetilde{\mathcal{W}}$ with annotated bounding boxes. Each object provides D" views" in an interlaced image. The interlaced bounding box is defined as the smallest bounding box that includes all bounding boxes of object "views". Fig. 7 illustrates this process.
- **Specialized Deep detector:** The proposed interlaced detector is a DCNN detector which allows to detect objects in the interlaced images.

  To do this, we specialize the Faster R-CNN deep detector [14] with a specific interlaced dataset $\tilde{\mathcal{W}}$.

  The specialization of the Faster R-CNN detector consists in adapting the parameters of the RPN network $\mathcal{R}$ for the object localization and the Fast R-CNN network $\mathcal{F}$ for object classification. For the training of the RPN network, we utilize a sliding window strategy so as to produce N bounding boxes for every position on the feature map that is generated by the last convolutional layer, where every bounding box is centered on the sliding window and is associated with a scale and an aspect ratio. After that, we calculate the Intersection-over-Union (IoU) overlap between the boxes of the interlaced dataset $\tilde{\mathcal{W}}$ and the bounding boxes generated by the sliding window with different ratios and scales. A proposal will be designated as a negative example if its maximum IoU with any box of the interlaced dataset $\tilde{\mathcal{W}}$ is less than another predefined threshold γ. A proposal will be designated as a positive training example if it overlaps with the interlaced dataset $\tilde{\mathcal{W}}$ box having an IoU greater than a predefined threshold λ, or if it is the bounding box that has the highest IoU with $\tilde{\mathcal{W}}$. After training the RPN, these proposals are used to train the Fast R-CNN. Therefore, a new specialized RPN ($\tilde{\mathcal{R}}$) and Fast R-CNN ($\tilde{\mathcal{F}}$) networks are generated after training with the interlaced dataset. These networks generate a new interlaced deep detector for the interlaced video, according to Eq. (13):

$$\left\{\tilde{\mathcal{R}}, \tilde{\mathcal{F}}\right\} = f_t\left(\tilde{\mathbf{I}}, \tilde{\mathcal{W}}; \mathcal{R}, \mathcal{F}\right) \tag{13}$$

where f_t is the training function, $\tilde{\mathcal{W}}$ is the interlacing dataset that contains the interlaced bounding boxes and $\tilde{\mathbf{I}}$ represents the interlaced set of images (see Eq. (1)).

4. Experimentation

This section presents the various tests performed to evaluate the performance of our multi-object tracking framework with the relevant ones on several public datasets and recent tracking benchmarks.

4.1. Evaluation datasets

Our framework has been evaluated on publicly available benchmarks and datasets: The MOT17 benchmark [46], the PETS 2009 [47] and the TUD [48] datasets. These benchmarks and datasets are mainly differentiated in terms of the number of tracking objects and fields of views. Fig. 8 shows example images of the evaluated datasets. The details are listed as follows.

- **PETS 2009:** The PETS 2009 dataset shows an outdoor scene where numerous people enter, interact, and exit with each other frequently. The major challenges of this dataset are frequent occlusions either caused by pedestrian interaction. Additionally

Fig. 8. Image examples of evaluated datasets: (a), (b): Images from the MOT17 benchmark. (c), (d) from the PETS dataset. (e), (f): Images from the TUD dataset.

Table 1
MOTA comparison for several interlacing configurations on several sequences of TUD public dataset. The best configuration is $D = 2, s = 2, g = 1$.

Sequence	$(D = 2, s = 2, g = 1)$	(2,2,2)	(2,8,1)	(4,2,1)	(2,2,4)	(1,1,1)
TUD-Stadtmitte	92.8%	**93.5%**	69.3%	85.5%	87.5%	87.5%
TUD-Campus	**88.1%**	64%	45.4%	70.8%	–%	79%
TUD-Crossing	**84.8%**	69.2%	33.1%	65.5%	–%	84.4%

Table 2
Table summarizing results of our framework on PETS and TUD sequences. Bold indicates best value for each column for each dataset. Abbreviations are as follows: MT - Mostly Tracked, ML - Mostly Lost, IDs - ID swaps, FM - Fragmentation.

PETS and TUD benchmarks						
Sequence	Method	MOTA	MT	ML	FM	IDs
S2L1	MHT-DAM (28)	83.5%	18	0	35	25
	DO (17)	86%	–	0	–	–
	NF (34)	85.5%	18	0	74	56
	Milan (56)	90.3%	18	0	52	22
	Li (55)	86.2%	15	0	–	11
	Ju (54)	75.9%	–	0	–	3
	FRCNN-MHT	87.4%	19	0	71	15
	FRCNNVI-MHT	**91.0 %/3.6%**	18	0	55	11
S2L2	MHT-DAM (28)	50.2%	7	2	207	197
	DO (17)	68%	–	–	–	–
	NF (34)	50.4%	6	3	379	244
	Milan (56)	58.1%	11	1	153	167
	Li (55)	51.5%	11	3	–	154
	Ju (54)	**71.6%**	37	0	–	225
	FRCNN-MHT	57.8%	20	0	305	268
	FRCNNVI-MHT	63.9%/ **6.1%**	24	0	280	232
S2L3	MHT-DAM (28)	35.6%	8	23	34	45
	NF (34)	**40.3%**	12	17	50	44
	Milan (56)	39%	8	19	22	27
	FRCNN-MHT	28.6%	9	6	188	204
	FRCNNVI-MHT	32.7%/**4.1%**	6	18	147	138
TUD-Stadtmitte	MHT-DAM (28)	61.4%	4	0	13	19
	NF (34)	51.6%	2	0	22	15
	Milan (56)	56.2 %	4	0	13	15
	FRCNN-MHT	87.5%	9	0	6	7
	FRCNNVI-MHT	**92.8%/5.3%**	8	0	4	3
TUD-Campus	FRCNN-MHT	79%	5	0	6	2
	FRCNNVI-MHT	**88.1%/9.1%**	7	0	4	0

to the widely used S2L1 and S2L2 sequence, we also evaluate our framework on the challenging S2L3 sequence that shows much denser crowds. The reason to use the S2L1, S2L2, and S2L3 sequences from the PETS 2009 dataset, respectively, corresponds to three representative application scenarios of MOT with low, high, and crowded object densities.

- **TUD-Stadtmitte** sequence is a video captured by a static camera at about a 2-meter height. This sequence shows walking people on the street.
- **TUD-Campus** sequence is a short video scene with frequent overlapping persons walking along a side walk. This sequence present large changes in scale and strong occlusions.
- **MOT17 benchmark:** we tested our tracking framework on the MOT17 benchmark [46] and we achieved very competitive results. There are seven training sequences and seven test sequences in the benchmark.

We demonstrate the evaluation results on testing sequences in order to verify the effectiveness of our framework.

4.2. Implementation details

In what follows, we will describe the implementation details of the proposed tracking framework.

As mentioned in Section 3.2, the detection step is achieved by the specialized Faster R-CNN deep detector [49]. We use the pre-trained VGG16 deep model [50] to initialize the Faster R-CNN, which has been used in several state-of-the-art detection approaches [51,14].

Following multiple experiments, we utilize the next parameters: 9 bounding boxes with 3 aspect ratios [1:2, 2:1, 1:1] and 3 scales $[512^2, 256^2, 128^2]$ produced on each position of the sliding window. At the training stage, 0.7 is the threshold λ of the IoU to select the positive samples and 0.3 is the threshold γ for the negative ones to build the training dataset.

The specialization of the Faster R-CNN is done by using annotated interlaced images built from several public training datasets from ETH-Bahnhof sequence [52], PETS2009 training set [47] and TUD-crossing sequence [48].

The association between detected interlaced bounding boxes is performed by the revisited version of MHT [28].

The latter has shown a substantial performance gain over the old versions of MHT by including both an online appearance and spatio-temporal models for multiple hypothesis tracking. The suggested

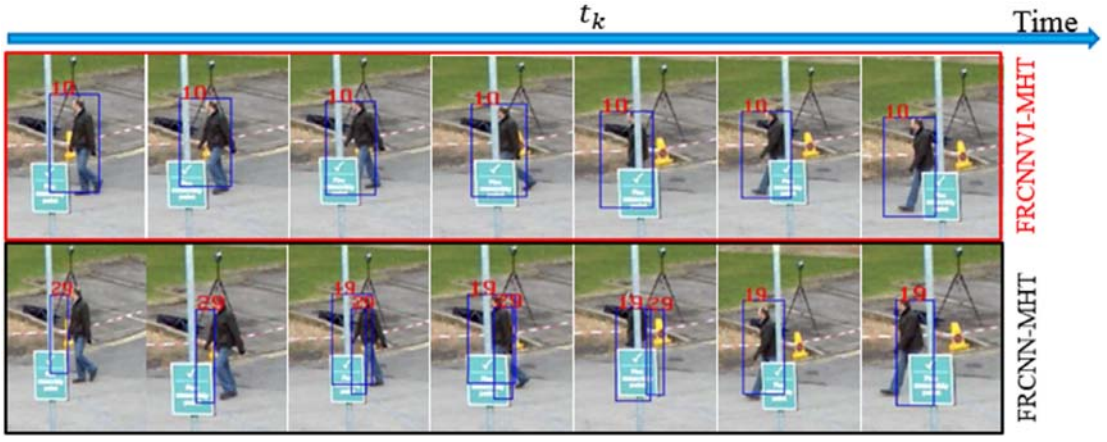

Fig. 9. Comparison between proposed interlaced MOT framework (FRCNNVI-MHT for top row images) and baseline one (FRCNN-MHT for bottom row images) for two challenging situations: occlusion and pedestrian crossing.

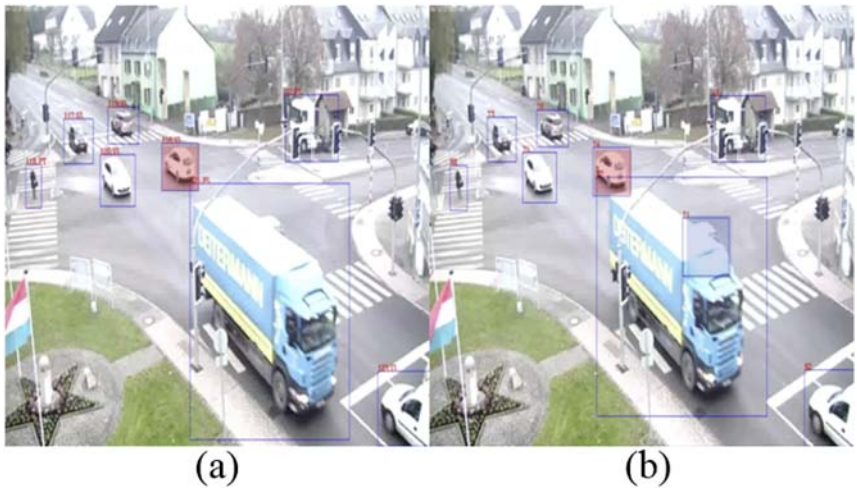

Fig. 10. Efficiency of our tracking framework to solve occlusion problems. (a) The truck hides the car and the latter is no longer detected and tracked. (b) Car (blue blob) is still detected and its identifier is preserved due to our interlacing strategy. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

framework has been tested with this association algorithm and we only use the spatio-temporal model. However, our framework is totally generic and it can be tested with other association algorithms.

4.3. Evaluation metrics

Performance evaluation is achieved using CLEARMOT metrics defined for visual multi-target tracking and detailed in [53]. The following metrics are taken into account: the MOT Accuracy (MOTA), the multiple MOT Precision (MOTP), the number of identity changes (IDS), the number of False Positives (FP) and the number of Missing Positions (MS). The MOTP considers only the localization precision of individuals without taking into account identity changes. The MOTA is a score which takes into consideration false negatives, false positives and identity switches of output trajectories. The MOTA metric is considered as the most important metric to evaluate the quality of the tracking.

4.4. Description of experiments

The proposed tracking framework (FRCNNVI-MHT) is compared with several state-of-the-art MOT frameworks:

- FRCNN-MHT: It is a Faster R-CNN detector and an revisited MHT (MHT-DAM) association method without interlacing step. This is the baseline for our comparison. It is important to note that the same sequences are used to train the baseline FRCNN-MHT and the proposed FRCNNVI-MHT.
- DO (2017) [17]: A tracking method was proposed for multi-object tracking with a tracklet association algorithm.
- Ju (2017) [54]: A novel multi-object tracking method based on frame-by-frame association with a novel affinity and appearance models.
- Li (2017) [55]: A tracking framework based on a novel fuzzy data association algorithm, which can significantly lead a better tracking result in visual multi-object tracking.

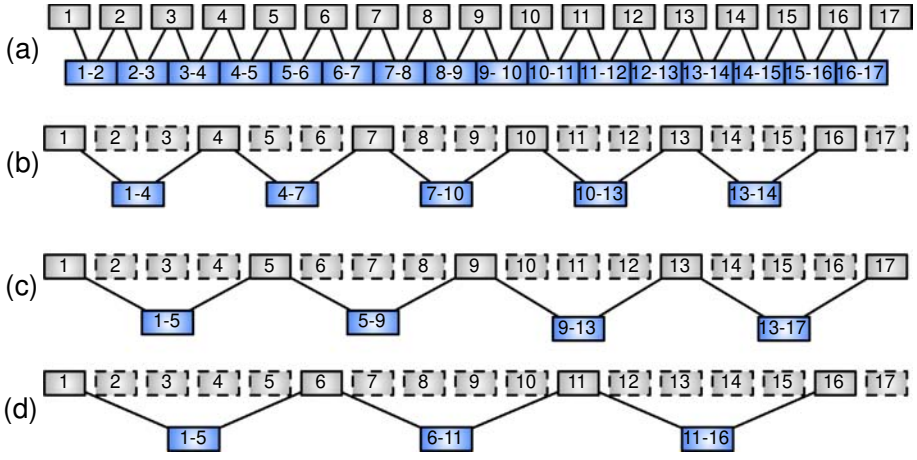

Fig. 11. Frame skipping strategy with four different sets of parameters for interlaced model: (a) $D = 2, s = 2, g = 1$, (b) $D = 2, s = 6, g = 3$, (c) $D = 2, s = 8, g = 4$ and $D = 2, s = 10, g = 5$. The top row with gray images represents the original video sequence, and the bottom row with blue images represents the interlaced video sequence. These images are never processed (skipped images). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

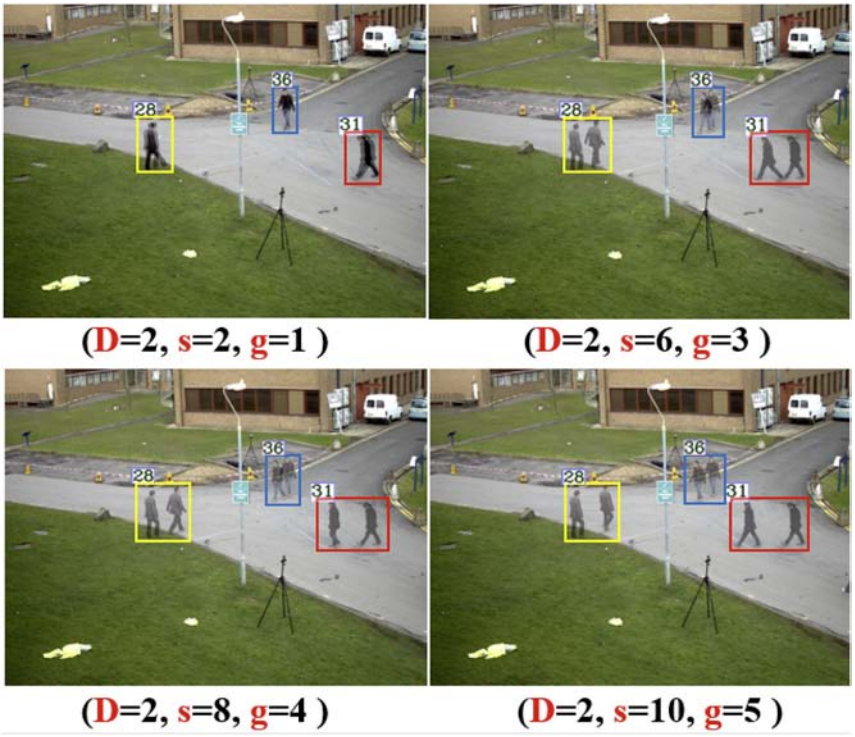

Fig. 12. Output examples of the interlaced specialized object detector for four interlacing strategies.

- NF (2016) [34]: A tracking framework suggested a pairwise cost to enforce tracklets, which effectively handled overlapping problems and tracking enhancements. The detection set is provided by the MOT benchmark.
- MHT-DAM (2015) [28]: This is the same association algorithm that we use in FRCNNVI-MHT and FRCNN-MHT but the detections are given by the MOT benchmark and an appearance model is combined with the spatio-temporal one.
- Milan (2013) [56]: A tracking framework was suggested to formulate the tracking problem by first selecting tracklets and then connecting them using a learned conditional random field. The detection set is provided by the MOT benchmark.

4.5. Results and analysis

This section presents the experiments realized in order to show the performances of the proposed video-interlacing model for the MOT. After testing several sets of parameters for the video-interlacing model, we compare the MOT framework with the baseline method and state-of-the-art tracking frameworks. The last experiment indicates that video-interlacing model can be used to reduce the computation time using a set of parameters to produce an image skipping strategy.

Since the main objective of the suggested video-interlacing model is to increase MOT performances, a first experiment is proposed to compare, for several sets of interlaced parameters, the benefit of our contribution. Table 1 presents the MOTA evaluation metric for several selected configurations. Results show that if some configurations improve the MOTA comparing with the baseline method (last column), others provide weak performance. Best results are obtained with the configuration: $D = 2, s = 2,$ and $g = 1$. This configuration will be set by default.

Table 2 shows the performances of the proposed framework "FRCNNVI-MHT" compared with the baseline "FRCNN-MHT" (without interlacing) and state-of-the-art tracking frameworks. For a fair comparison, the ground truth annotations and the evaluation script provided by Milan et al. [56] are used. The MOTA increases for all tested datasets.

The red value on the last line of each row of Table 2 represents the improvement of our interlacing framework over the baseline one. The median improvement is **6%** in all evaluation datasets. Fig. 9 illustrates a comparison between our proposed framework (FRCNNVI-MHT) and the baseline one (FRCNN-MHT) on two challenging situations: occlusion and pedestrian crossing.

Compared to the state-of-the-art tracking frameworks on the PETS2009 and the TUD sequences, the Faster R-CNN combined with the MHT-DAM association algorithm [28] is very competitive. Our framework has significant improvements and enhances the result of the MHT-DAM algorithm by about **12.5%** on average as a MOTA evaluation metric (**31.4%** for TUD-Stadtmitte sequence). However, for sequence S2L3, which represents a dense crowd, the Faster R-CNN gives poor results compared to the detection set provided by the MOT benchmark (MHT-DAM). Fig. 10 demonstrates a comparison over the proposed framework FRCNNVI-MHT and the baseline FRCNN-MHT one on a private vehicle dataset, in order to show the efficiency of the suggested interlacing model for multi-object tracking (car, person, truck, ...) and the robustness to solve the occlusion/intersection problems in case when one vehicle is masked by another.

Fig. 11 illustrates four interesting interlacing configurations that produce frame skipping; i.e., for ($D = 2, s = 6, g = 3$), odd images are never processed, resulting in a reduced computation time and for

Table 3
MOTA evaluation metric for several interlacing configurations selected to produce frame skipping on TUD dataset. The performance results using the interlaced model are presented in black color, and without the interlaced model (but with frame skipping) are in blue.

Interlacing configurations					
Sequence	Method	(D=2,s=2,g=1)	(2, 6, 3)	(2, 8, 4)	(2, 10, 5)
TUD-Stadtmitte	FRCNNVI	**92.8%** |**87.5%**	92.7% | –%	79.8% | –%	74% | –%
TUD-Campus	FRCNNVI	**88.1%** |**79%**	61.7% | –%	30.8% | –%	20.4% |–%
TUD-Crossing	FRCNNVI	**84.8%** |**84.4%**	36.6% |–%	30.3% | –%	21.7% |–%

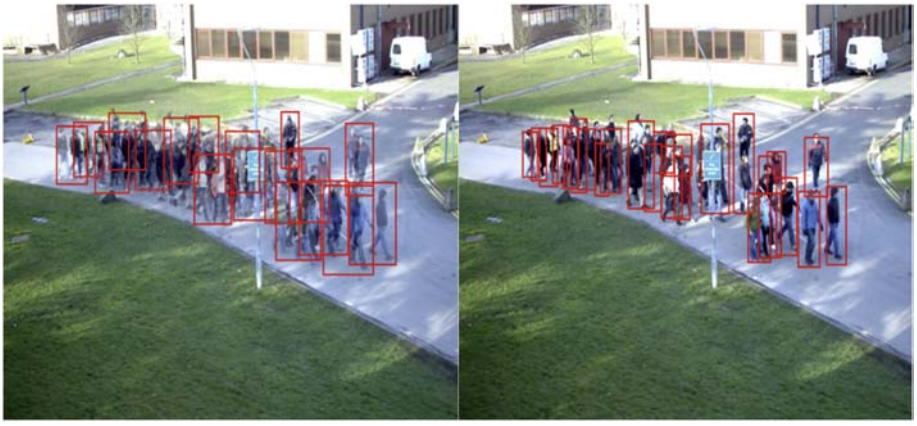

Fig. 13. Examples of FRCNN-MHT failures (with interlacing configuration ($D = 2, s = 6, g = 3$)). The reason of that is due to wrong detections provided by specialized Faster R-CNN detector: The left image shows detection results provided by specialized interlaced detector and the right one for Faster R-CNN. In dense crowd, the interlaced video mixes pedestrians resulting to a limitation of the method.

($D = 2, s = 10, g = 5$), **25%** of the images are processed. Fig. 12 presents examples of the output of the interlaced specialized object detector for the four interlacing configurations mentioned above.

The results are provided in Table 3. The MOTA for the proposed framework (FRCNNVI-MHT) is represented in black while the MOTA for the baseline one (FRCNN-MHT) applied with frame skipping is represented in blue. Please notice that the MHT-DAM association fails from two skipped images. The main reason is that our implementation of the MHT-DAM uses only a spatio-temporal model for tracking, which will fail when there is no overlap between detection. The proposed interlacing framework produces overlapped detection, which improves the performances of the association.

As mentioned in Table 3, for TUD-Stadmitte, the MOTA performance for four skipped frames decreases by about 10% compared to the non skipped frames strategy. Since video-interlacing and reverse video-interlacing are very low CPU time consuming related to the detector, skipping frame strategies provides an efficient way to decrease the computation time of the MOT while maintaining competitive performances.

However, the FRCNN-MHT applied with frame skipping gives bad results (as shown by–% in Table 3) due to the limitation of the Faster R-CNN detector to detect objects on crowded sequences (see Fig. 13).

4.6. MOT17 benchmark comparison

To illustrate the effectiveness of our approach, we also perform our tracking experiments on the MOT17 Benchmark. The test set contains 7 sequences, where imaging condition, camera angle, and camera motion are very different. For each test sequence, the benchmark also provides a training sequence that is captured in the similar setting. The specialization of the Faster R-CNN is done by using annotated interlaced images built from the 7 training sequence provided by the MOT17 benchmark. The final result of testing sequences provided by our proposed framework is submitted to the benchmark.

Fig. 14 presents a comparison between the baseline FRCNN-MHT (Faster R-CNN & MHT-DAM without interlacing) and our proposed one FRCNNVI-MHT in all MOT17 test sequences. Our framework has significant improvements and enhances the result of the baseline framework by about **9.3%** on average as a MOTA evaluation metric on MOT17 benchmark. It proves our effectiveness of our interlacing model.

In Fig. 14, it is noted that on sequence "MOT17-07" of the MOT17 benchmark, we have only 1% improvement which is explained by the low complexity of the scene, so we cannot really see the advantages of interlacing in solving multi-object tracking issues.

5. Conclusion

In this paper, we have presented a new MOT-bD framework which proposes to built an intermediate interlaced video-sequence and an associated DCNN detector. The resulting MOT-bD algorithm improves the tracking performance in complex videos taken by a mobile or stationary cameras. Our suggested tracking framework is generic and can be used with other association algorithms. Moreover,

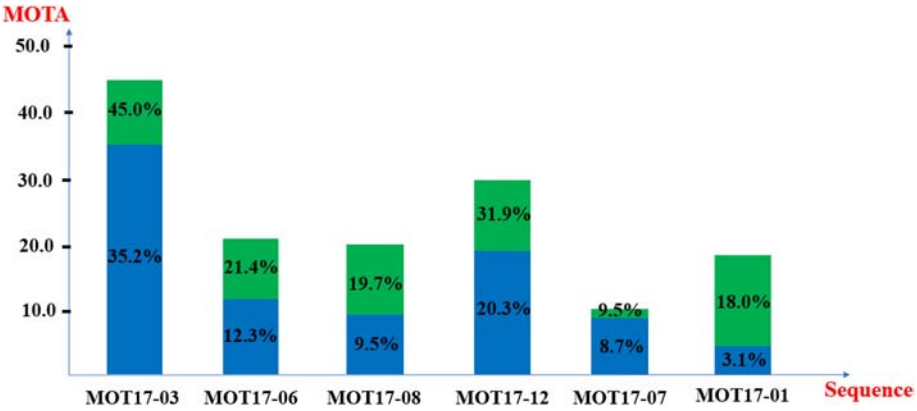

Fig. 14. Summarizing results of our framework on MOT17 benchmark. The blue color in the each stacked bar presents the result of the baseline FRCNN-MHT (without interlacing) and the green one presents the improvement given by our FRCNNVI-MHT on each test sequence of the MOT17 benchmark. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

we have demonstrated that some interlacing strategies can be proposed to skip frames and reduce complexity during tracking, while maintaining a good performance.

The proposed framework implies that annotated video sequences have to be available to train the specialized interlaced DCNN. Future work will focus in automatic specialization using domain adaptation algorithms for the MOT-bD.

Acknowledgments

This work is within the scope of a co-guardianship between the University of Sousse (Tunisia) and Clermont Auvergne University (France). It is sponsored by the French government research program “Investissements d’avenir” through the IMobS3 Laboratory of Excellence (ANR-10-LABX-16-01), by the European Union through the program Regional competitiveness and employment 2007–2013 (ERDF - Auvergne region), and by the Tunisian Ministry of Higher Education & Scientific Research.

Appendix A. Supplementary data

You will find some video demonstrations attached with the manuscript which illustrate the robustness and the "power of video-interlacing to improve the MOT-bD. Supplementary data to this article can be found online at https://doi.org/10.1016/j.imavis.2019.03.002.

References

[1] A. Mhalla, T. Chateau, S. Gazzah, N.E.B. Amara, An embedded computer-vision system for multi-object detection in traffic surveillance, IEEE Transactions on Intelligent Transportation Systems (2018)
[2] A. Mhalla, T. Chateau, H. Maamatou, S. Gazzah, N.E.B. Amara, SMC faster R-CNN: toward a scene-specialized multi-object detector, Computer Vision and Image Understanding 164 (2017) 3–15.
[3] A. Milan, S. Roth, K. Schindler, Continuous energy minimization for multitarget tracking, IEEE transactions on pattern analysis and machine intelligence 36 (1) (2014) 58–72.
[4] X. Wang, E. Türetken, F. Fleuret, P. Fua, Tracking interacting objects optimally using integer programming, European Conference on Computer Vision, Springer. 2014, pp. 17–32.
[5] B. Han, J. Sim, H. Adam, BranchOut: regularization for online ensemble tracking with convolutional neural networks, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[6] M. Danelljan, G. Häger, F.S. Khan, M. Felsberg, Discriminative scale space tracking, IEEE transactions on pattern analysis and machine intelligence 39 (8) (2017) 1561–1575.
[7] S.-H. Bae, K.-J. Yoon, Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning, Proceedings of the IEEE conference on computer vision and pattern recognition, 2014. pp. 1218–1225.
[8] A. Dehghan, M. Shah, Binary quadratic programing for online tracking of hundreds of people in extremely crowded scenes, IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)
[9] Z. Shen, Z. Liu, J. Li, Y.-G. Jiang, Y. Chen, X. Xue, DSOD: learning deeply supervised object detectors from scratch, The IEEE International Conference on Computer Vision (ICCV), 2017.
[10] E. Kalogerakis, M. Averkiou, S. Maji, S. Chaudhuri, 3D shape segmentation with projective convolutional networks, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[11] S. Rota Bulo, G. Neuhold, P. Kontschieder, Loss max-pooling for semantic image segmentation, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[12] L. Wang, W. Ouyang, X. Wang, H. Lu, Stct: sequentially training convolutional networks for visual tracking, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. pp. 1373–1381.
[13] M. Wang, Y. Liu, Z. Huang, Large margin object tracking with circulant feature maps, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[14] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, Advances in neural information processing systems, 2015. pp. 91–99.
[15] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, Ssd: single shot multibox detector, European conference on computer vision, Springer. 2016, pp. 21–37.
[16] J. Redmon, A. Farhadi, YOLOv3: an incremental improvement, 2018.arXiv.
[17] Y. Dorai, F. Chausse, S. Gazzah, N.E.B. Amara, Multi target tracking by linking tracklets with a convolutional neural network, VISIGRAPP (6: VISAPP), 2017. pp. 492–498.
[18] S. Tang, M. Andriluka, B. Andres, B. Schiele, Multiple people tracking by lifted multicut and person re-identification, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. pp. 3539–3548.
[19] B. Wang, L. Wang, B. Shuai, Z. Zuo, T. Liu, K. Luk Chan, G. Wang, Joint learning of convolutional neural networks and temporally constrained metrics for tracklet association, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016. pp. 1–8.
[20] G. Shu, A. Dehghan, O. Oreifej, E. Hand, M. Shah, Part-based multiple-person tracking with partial occlusion handling, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012. pp. 1815–1821.
[21] A. Milan, S.H. Rezatofighi, A.R. Dick, I.D. Reid, K. Schindler, Online multi-target tracking using recurrent neural networks., AAAI, 2017. pp. 4225–4232.
[22] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, Proceedings of the IEEE conference on computer vision and pattern recognition, 2015. pp. 1–9.
[23] W. Feng, Z. Hu, W. Wu, J. Yan, W. Ouyang, Multi-object tracking with multiple cues and switcher-aware classification,arXiv preprint arXiv:1901.06129. 2019.
[24] R. Henschel, L. Leal-Taixé, D. Cremers, B. Rosenhahn, Fusion of head and full-body detectors for multi-object tracking, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE. 2018, pp. 1509–150909.
[25] H. Yoo, K. Kim, M. Byeon, Y. Jeon, J.Y. Choi, Online scheme for multiple camera multiple target tracking based on multiple hypothesis tracking, IEEE Transactions on Circuits and Systems for Video Technology 27 (3) (2017) 454–469.
[26] B. Lee, J. Park, Y. Joo, S. Jin, Intelligent Kalman filter for tracking a manoeuvring target, IEE Proceedings-Radar, Sonar and Navigation 151 (6) (2004) 344–350.
[27] J. Vermaak, S.J. Godsill, P. Perez, Monte Carlo filtering for multi target tracking and data association, IEEE Transactions on Aerospace and Electronic systems 41 (1) (2005) 309–332.
[28] C. Kim, F. Li, A. Ciptadi, J.M. Rehg, Multiple hypothesis tracking revisited, Proceedings of the IEEE International Conference on Computer Vision, 2015. pp. 4696–4704.
[29] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, P.H.S. Torr, End–to-end representation learning for correlation filter based tracking, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[30] J. Badie, F. Bremond, Global tracker: an online evaluation framework to improve tracking quality, Advanced Video and Signal Based Surveillance (AVSS), 2014 11th IEEE International Conference on, IEEE. 2014, pp. 25–30.
[31] I.J. Cox, S.L. Hingorani, An efficient implementation of Reid's multiple hypothesis tracking algorithm and its evaluation for the purpose of visual tracking, IEEE Transactions on pattern analysis and machine intelligence 18 (2) (1996) 138–150.
[32] M. Han, W. Xu, H. Tao, Y. Gong, An algorithm for multiple object trajectory tracking, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1, IEEE. 2004, pp. I–I.
[33] M. Danelljan, F. Shahbaz Khan, M. Felsberg, J. Van de Weijer, Adaptive color attributes for real-time visual tracking, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014. pp. 1090–1097.
[34] V. Chari, S. Lacoste-Julien, I. Laptev, J. Sivic, On pairwise costs for network flow multi-object tracking, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015. pp. 5537–5545.
[35] N. McLaughlin, J.M. Del Rincon, P. Miller, Enhancing linear programming with motion modeling for multi-target tracking, Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on, IEEE. 2015, pp. 71–77.
[36] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, Y. Wu, Learning fine-grained image similarity with deep ranking, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014. pp. 1386–1393.
[37] E. Hoffer, N. Ailon, Deep metric learning using triplet network, International Workshop on Similarity-Based Pattern Recognition, Springer. 2015, pp. 84–92.
[38] S. Chopra, R. Hadsell, Y. LeCun, Learning a similarity metric discriminatively, with application to face verification, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1, IEEE. 2005, pp. 539–546.
[39] Y. Taigman, M. Yang, M.A. Ranzato, L. Wolf, Deepface: closing the gap to human-level performance in face verification, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014. pp. 1701–1708.
[40] Y. Sun, Y. Chen, X. Wang, X. Tang, Deep learning face representation by joint identification-verification, Advances in neural information processing systems, 2014. pp. 1988–1996.
[41] R. Tao, E. Gavves, A.W. Smeulders, Siamese instance search for tracking, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. pp. 1420–1429.
[42] L. Leal-Taixé, C. Canton-Ferrer, K. Schindler, Learning by tracking: Siamese CNN for robust target association, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016. pp. 33–40.
[43] B. Kumar, G. Carneiro, I. Reid, Learning local image descriptors with deep Siamese and triplet convolutional networks by minimising global loss functions, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. pp. 5385–5394.
[44] D. Cheng, Y. Gong, S. Zhou, J. Wang, N. Zheng, Person re-identification by multi-channel parts-based CNN with improved triplet loss function, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. pp. 1335–1344.

[45] S. Tang, M. Andriluka, B. Schiele, Detection and tracking of occluded people, International Journal of Computer Vision 110 (1) (2014) 58–69.
[46] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, K. Schindler, MOT16: a benchmark for multi-object tracking,arXiv preprint arXiv:1603.00831. 2016.
[47] J. Ferryman, A. Shahrokni, Pets2009: Dataset and challenge, Performance Evaluation of Tracking and Surveillance (PETS-Winter), 2009 Twelfth IEEE International Workshop on, IEEE. 2009, pp. 1–6.
[48] M. Andriluka, S. Roth, B. Schiele, People-tracking-by-detection and people-detection-by-tracking, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE. 2008, pp. 1–8.
[49] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, Advances in neural information processing systems, 2015. pp. 91–99.
[50] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition,arXiv preprint arXiv:1409.1556. 2014.
[51] R. Girshick, Fast r-cnn, Proceedings of the IEEE International Conference on Computer Vision, 2015. pp. 1440–1448.
[52] A. Ess, B. Leibe, K. Schindler, L. van Gool, A mobile vision system for robust multi-person tracking, IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08), IEEE Press. 2008,
[53] K. Bernardin, R. Stiefelhagen, Evaluating multiple object tracking performance: the CLEAR MOT metrics, EURASIP Journal on Image and Video Processing 2008 (1) (2008) 246309.
[54] J. Ju, D. Kim, B. Ku, D.K. Han, H. Ko, Online multi-object tracking with efficient track drift and fragmentation handling, JOSA A 34 (2) (2017) 280–293.
[55] L.-Q. Li, E.-Q. Li, W.-M. He, A novel fuzzy data association approach for visual multi-object tracking, ITM Web of Conferences, 12, EDP Sciences. 2017, pp. 05004.
[56] A. Milan, K. Schindler, S. Roth, Detection-and trajectory-level exclusion in multiple object tracking, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013. pp. 3682–3689.