Online Appearance-Motion Coupling for Multi-Person Tracking in Videos

Bonan Cuan, Khalid Idrissi, and Christophe Garcia

***Abstract*. Multi-person tracking in videos is a promising but challenging visual task. Recent progress in this field has introduced deep convolutional features as appearance models, which achieve robust tracking results when coupled with proper motion models. However, model failures that often cause severe tracking problems have not been well discussed and addressed in previous work. In this paper, we propose a solution using online detection of such failures and accordingly adjusting the coupling between appearance and motion models. The strategy is to let the functional models take over when certain models face data association ambiguity and simultaneously suppress the influence of inappropriate observations during the model update. Experimental results have proven the benefit of our proposed improvement.**

***Index Terms*. Multiple object tracking, deep neural network, online learning, tracking-by-detection, multiple hypothesis tracking.**

I. Introduction

Tracking people in videos is a computer vision task with many practical applications (e.g., video surveillance, autonomous driving, and human-computer interface). As a multi-object tracking (MOT) problem, it has several intrinsic challenges compared to generic object tracking tasks. The non-rigidity of people requires robust trackers to cope with person pose variation. The situation deteriorates when tracking a crowd simultaneously. People in the scene often interact with each other or with other objects, which incurs more occlusions, scene clutter, and complex object dynamics.

Owing to the deep neural networks [1]-[3] widely applied in object recognition, multi-object tracking has witnessed significant progress under the paradigm of "tracking-by-detection." Objects are detected in each frame and are then associated into trajectories through frames, during which their appearance and motion information serves as important guidance of track inference [4]-[6].

On the one hand, appearance models, especially those based on deep learning techniques [5], [6], have proven to be more robust for not only object detection but also intraclass recognition. Either a discriminant convolutional representation [5] or the entire classifier [6] is learned to identify detected objects. With their help, object re-identification after long occlusions becomes reliable.

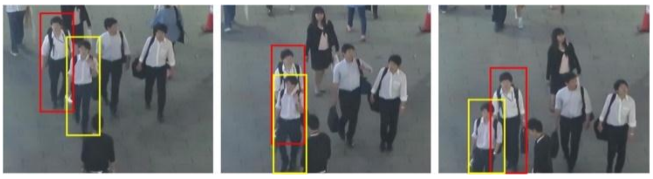


Fig. 1. Example of adjacent similar-looking people in MOT16 benchmark [7]. They are difficult for appearance models to distinguish. Once their trajectories cross, motion models are more robust for preventing mismatches or trajectory merging.

On the other hand, despite the remarkable growth of appearance models, similar looking people may be indistinguishable only according to their appearances, even by human experts. Motion models are indispensable in tracking. An example of a Multiple Object Tracking Challenge 2016 (MOT16) [7] is shown in Fig. 1. From the simplest neighbor gating [4] to more complex ones like linear motion models, linear quadratic estimation with a Kalman filter [5] or spatiotemporal relation metric [6], motion models help reduce search space for data association and to boost object re-identification.

Given their own shortcomings, the coupling of both types of models is important for good tracking. For most state-of-the-art tracking algorithms, the coupling is as simple as combining all models with pre-chosen coefficients controlling their relative weights. An association decision is made based on such weak couplings, while after each established association, every model is updated nearly independently to incorporate new instances. The coupling weights are determined offline and remain invariant during the entire tracking process.

Nevertheless, such a strategy has drawbacks. Firstly, when a model fails (e.g., an appearance model itself has trouble differentiating adjacent similar people in Fig. 1), it should be assigned with a lower weight to let the other model or models take over the decision making of data association. Secondly, during the model update, false positive instances caused by the failure of one model could be absorbed into the other models.

Therefore, in this paper, we come up with an online appearance-motion coupling approach. The weights of models during both decision making and model update phases are calculated online, according to their credibility in each step. Local apparent ambiguity means entrusting the motion model more with the decision making, and vice versa. Our main contributions will be detailed in Section III after a brief review of the related work. Experimental results that prove the effectiveness of our proposed model will be shown and discussed in Section IV. The conclusion and perspectives will be found at the end of this paper.

Manuscript received February 24, 2019; revised April 12, 2019. This work was supported in part by the Chinese Scholarship Council.

The authors are with the Universit éde Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205, F-69621 France (e-mails: bonan.cuan@insa-lyon.fr, khalid.idrissi@insa-lyon.fr, christophe.garcia@insa-lyon.fr).

DOI: 10.7763/IJMO.2019.V9.687

II. Related Work

Recent work in multi-person tracking is mostly driven by MOT challenges [7] and focuses on tracking-by-detection algorithms. The benchmark provides public detection results shared by all submitted algorithms. The effectiveness and robustness of data association methods are the key points to inspect.

The trend is to combine deep convolutional features with well-designed motion models. Kim *et al.* [5] introduced convolutional features as an appearance model into the MHT framework and solved the search space explosion problem. Discriminant features are employed to efficiently trim the hypothesis trees which can experience high combinatoric growth when pruned only with the Gaussian estimation motion model and unreliable appearance models. Another exemplary success is [6]. Tang *et al.*'s model multi-person tracking is an offline lifted multi-cut problem (LMP). Such a graph-based tracking algorithm aims to find the optimal cut of the quasi-complete multipartite graph that models objects in all frames. The cut is based on scores of a motion model (spatiotemporal relation metric) and two appearance models, including a deep stacked net classifier fusing pre-trained human body part detectors. With the help of robust appearance-based re-identification methods, the temporal scope of the graph cut was significantly enlarged.

However, both state-of-the-art algorithms did not take temporary failures of their appearance models into consideration. This paper argues the benefit of an online coupling of appearance and motion models and demonstrates the improvement after applying this strategy on MHT framework.

In this paper, we keep the original motion model in MHT, i.e., linear quadratic estimation with a Kalman filter, while adopting a deep Siamese network [8] as the appearance model. As a metric learning approach, the network extracts feature vectors of a pair of observations via the same deep convolutional backbone net. Features of the two branches are compared with each other under a Cosine Similarity metric [9]. The structure of the network will be briefly revisited in Section IV.

III. Online Model Coupling

The collaboration and interference of appearance and motion models mainly happens in two phases: data association relies on the combination of the models for making a track decision, after which all the models need to be accordingly updated. Our online coupling affects both phases, which are reported respectively in the following two subsections. The discussion is based on the MHT framework [5] yet can be readily extended to other algorithms.

Under the tracking-by-detection schema, tracking is done via linking newly detected objects (sometimes intermediate tracklets) with established track hypotheses in each frame. In the form of either graphcut [6] or tree growth [5], data association requires an association score S related to the possibility of a detection candidate belonging to a track hypothesis. The score S is often the output of a combination of appearance and motion models which are based on one single [6] or a bag of previous observations [4][5] in the hypothesis. Most tracking algorithms use the weighted sum of the scores, with the weights as pre-defined constants. We adapted the denotation from MHT [5]:

$$S = w_{mot}S_{mot} + w_{app}S_{app} \, , \qquad (1)$$

where $S_{mot/app}$ and $w_{mot/app}$ denote the output scores and their regularization weights of the motion and appearance models, respectively.

The decision making described in Eq. (1) focuses explicitly on whether a candidate observation o fits into a track hypothesis t . Nonetheless, data association is an injective problem from hypotheses $T = \{t_i\}$ to observations $O = \{o_j\}$: not only confirmed the track has a single observation (nor none) in each frame, but also an observation can be assigned to no more than one track. The latter part, or the observation's inverse selection of the hypothesis, was constantly ignored by the tracking algorithms. It was often implicitly achieved through concurrences among hypotheses whenever an observation has more than one plausible assignment: we define the feasible assignments of observation o as a subset of T :

$$T_o = \{t_i \mid S_{mot}(t_i,o) \ge \theta_{mot} \,\&\, S_{app}(t_i,o) \ge \theta_{app}\} \, , \qquad (2)$$

where θ_{mot} and θ_{app} are the trimming thresholds of the motion and appearance models, respectively. Ambiguity emerges when $|T_o| > 1$, with $|T_o|$ denoting the cardinality of the set. After the concurrence of suitable hypotheses, local mismatches (or ID switches) often come out (see Fig. 2 as an illustration).

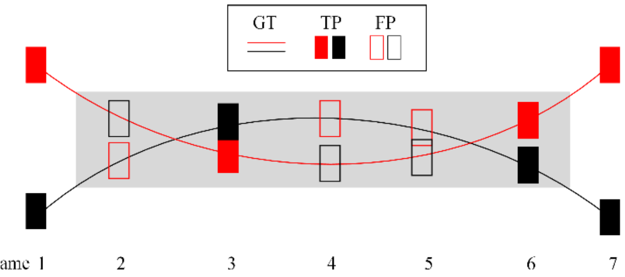


Fig. 2. Illustration of hypotheses concurrences. This happens when black and red trackers have comparable scores (often caused by an appearance model failure facing similar looking objects) for the same observation. The hypothesis-observation assignment can be unstable, which results in local mismatches and false positive trajectory parts. (In both Fig. 2 and 3, GT stands for groundtruth, TP for True Positive, FP for False Positive and Occ. for Occlusion.)

Hence, we introduce the concept of model credibility to assess if a model is reliable at a certain stage. Here, we propose an online method to determine the credibility of the appearance model of $t \in T_o$:

$$c_{app}(t,o) = \frac{e^{S_{app}(t,o)}}{\sum_{t \in T_o} e^{S_{app}(t,o)}} \qquad (3)$$

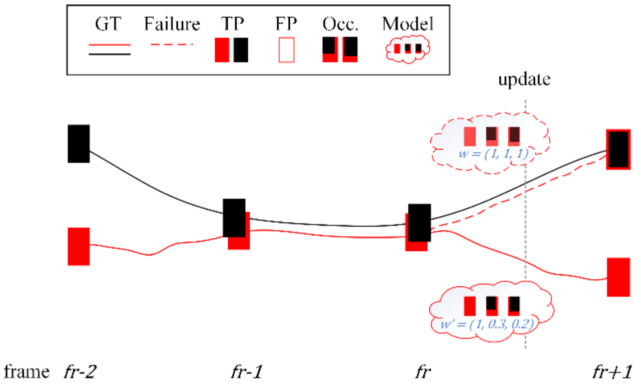


Fig. 3. Influence of noisy observations during the update of the model. Partially occluded observations (illustrated by overlapping bounding boxes) may cause inconsistency of the object appearance when given the same weight as benign instances (weight vector w). This may lead to invalid appearance models and finally track fragments or even trajectory merging (see dotted red lines). We force online noise pruning based on similarity scores, to minimize or eliminate the influence of outlier observations (the lower model with w').

We suggest the scores $S_{app}(t,o)$ should be normalized to keep their exponentiations comparable. Credibility $0<c_{app}<1$ describes how confident the appearance-based data association is. With the credibility, we define a variable coefficient c for model coupling:

$$c=\min(c_{app},\theta_c)\ , \tag{4}$$

in which $0<\theta_c<1$ is a threshold (e.g., $\theta_c=0.9$) to prevent the motion information from total disappearance in our new coupling function as shown below:

$$S'=(1-c)\cdot w_{mot}S_{mot}+c\cdot w_{app}S_{app} \tag{5}$$

It is designed to encourage *one-hot* appearance-based correspondence, which implies no noticeable ambiguity. The more evenly matched concurrences among the hypotheses exist, the more unreliable their appearance models are. Upon the failure of the appearance model, the motion model will step in for further arbitration.

The coefficient is calculated online using the hypotheses in the current frame and their appearance models. An online coupling based on the credibility of motion models may also be useful yet will not be discussed in the scope of this paper.

Online Appearance Model Update

After the decision making of the data association, each track hypothesis has a new object observation. The models need updates to incorporate such new instances. The update of the appearance models is much more important than that of the motion models, given that the appearance models are often more sophisticated and reliable over a long time interval. In this subsection, we deal with the problem occurring during the update of appearance models.

In MHT-based algorithm [5][8], each track hypothesis t in frame fr keeps a bag of feature vectors of previous observations $V=\{\vec{v}_{1,2,\cdots fr}\}$. The update of the appearance model is as simple as appending the feature vector $\vec{v}_{fr+1}$ of its new instance into its bag. All vectors in the bag are equally weighted. The updated model will guide the next step of the data association, to which every previous observation has the same contribution.

However, all to-be-absorbed observations are not always benign. Due to occlusions (especially those spanning long-duration and long distance) as well as abrupt object and/or camera movements, tracking algorithms often struggle to fill in the blank of missing targets by introducing ambiguous observations into the track, under the guidance of motion models. Partially occluded, imprecise or even negative observations of poor appearance scores are accepted only because of their advantageous locations. Appearance models can be contaminated by such noises once updated. The inconsistency in object appearance will then lead to mismatches, track fragments, and trajectory merging (see Fig. 3).

Therefore, appearance models need to be amended when motion models malfunction. In this paper, we inherit the appearance model in [8] which was originally represented by the average of all the normalized vectors in set V :

$$\vec{v}_{app}=\frac{\sum_{\vec{v}_k\in V}\frac{\vec{v}_k}{\|\vec{v}_k\|}}{|V|} \tag{6}$$

where $\|\bullet\|$ is the L2 norm of a vector, and the cardinality $|V|$ indicates the length of the track hypothesis t in the frame fr . Instead of entirely trusting a new observation fitting of the appearance model, we come up with an online sanity check.

Given the cosine similarity between its feature vector $\vec{v}_{fr+1}$ and the model,

$$sim(\vec{v}_{fr+1})=\vec{v}_{app}\cdot\frac{\vec{v}_{fr+1}}{\|\vec{v}_{fr+1}\|} \tag{7}$$

likelihood of an accurate appearance-based data association is defined by assuming the similarity scores follows a Gaussian distribution:

$$llh=\frac{f(sim;\mu_+,\sigma_+)}{f(sim;\mu_+,\sigma_+)+f(sim;\mu_-,\sigma_-)} \tag{8}$$

where $f(sim;\mu,\sigma)$ is the probability density function of the Gaussian distribution $N(\mu,\sigma)$ with the mean μ and standard deviation σ . The statistical values of an accurate and inaccurate association are subscripted with $+$ and $-$, respectively. They are often recorded during the training phase of metric learning [8]. For the details of similarity metric learning in Eq. (7) and (8), please refer to [9][8].

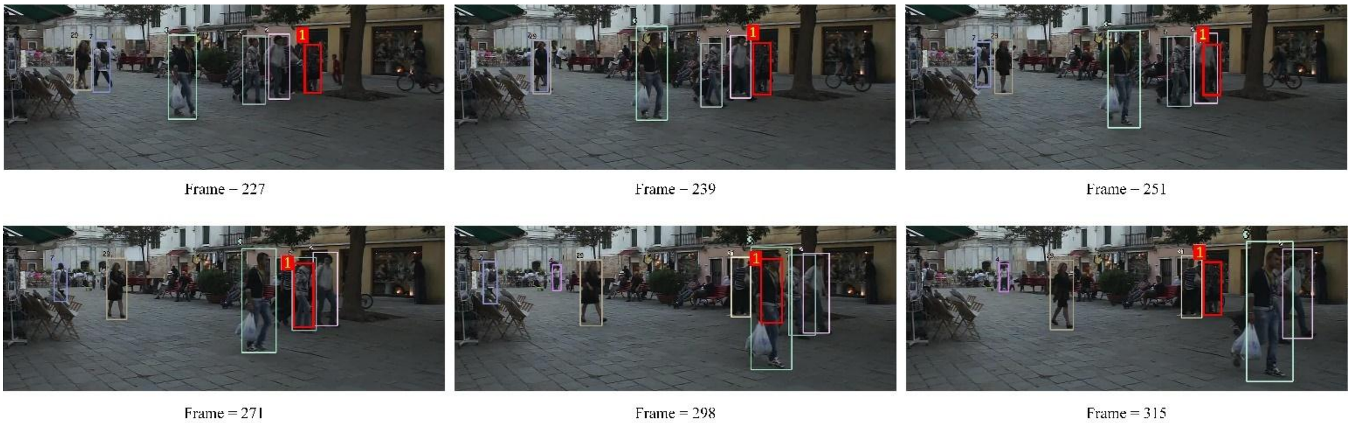
With the appearance-based association likelihood, the model update described in Eq. (6) can be improved: Every feature vector is weighted according to its similarity with the appearance model. Therefore, outliers have less influence on the object appearance. Besides this, the pairwise similarity in Eq. (7) and the likelihood in Eq. (8) are already calculated for the data association before the model update phase. The online update in Eq. (9) improvement has no extra overhead

$$\vec{v}_{app}\,' = \frac{\sum_{\vec{v}_k \in V} (llh_k \cdot \frac{\vec{v}_k}{\|\vec{v}_k\|})}{\sum_{\vec{v}_k \in V} llh_k} \tag{9}$$

TABLE I: MULTI-PERSON TRACKING RESULTS ON MOT16 BENCHMARK [7]

Method	**MOTA**	MOTP	MT	ML	FP	FN	IDsw	Frag
MHT_DAM v1 [5]	42.9	76.6	13.6%	46.9%	5668	97919	499	659
MHT_DAM v2 [5]	45.8	76.3	16.2%	43.2%	6412	91758	590	781
Baseline	44.3	76.0	14.9%	42.3%	9316	91331	883	850
Baseline + U	45.4	76.3	15.4%	43.2%	5735	92954	794	779
Baseline + U + C	**46.0**	76.3	15.5%	42.6%	5124	92697	693	759

MOT accuracy (MOTA) is the most important evaluation metric. For more instructions, please refer to the website of MOT Challenge.


Frame – 227

Frame – 239

Frame – 251

Frame = 271

Frame = 298

Frame = 315

Fig. 5. Samples of the tracking results of the sequence MOT16-01. With the proposed online appearance and motion models coupling, targets can be recovered after a long period of disappearance. For instance, the person with ID = 1 (red bounding box) is consistently tracked, even enduring an incessant occlusion of more than 60 frames.

IV. EXPERIMENTS

Experiments are conducted on the MOT16 benchmark [7] to demonstrate the effectiveness of our contributions. The deep Siamese network designed in [8] for metric learning is illustrated in Fig. 4. The realization specifications are inherited from [8] with some modifications: deep residual net with 50 layers [3] is chosen as the backbone for the convolutional feature extraction; it is trained only on the MOT16 training set, no other datasets are involved.

The tracking result of the original method in [8] (without any improvement described in section III) is set as the baseline. A test only with the online appearance update mechanism (section III. B) is then conducted (denoted as “baseline + U”). The last experiment includes both the contributions (denoted as “baseline + U + C”).

The evaluation metrics of the tracking results are listed in Table I, along with the original MHT algorithm (two submissions). Without abusive hyperparameter tuning, our algorithm outperforms the original MHT algorithm.

When compared to the baseline algorithm, the application of each of our online coupling strategies results in a notable improvement in most evaluation metrics. It proves the effectiveness of our online appearance-motion coupling. Besides this, no noticeable overhead is added given its simple structure and negligible calculation. On the contrary, hypothesis trimming becomes faster with the help of the proposed coupling model.

Some tracking results can be found in Fig. 5 at the end of the paper.

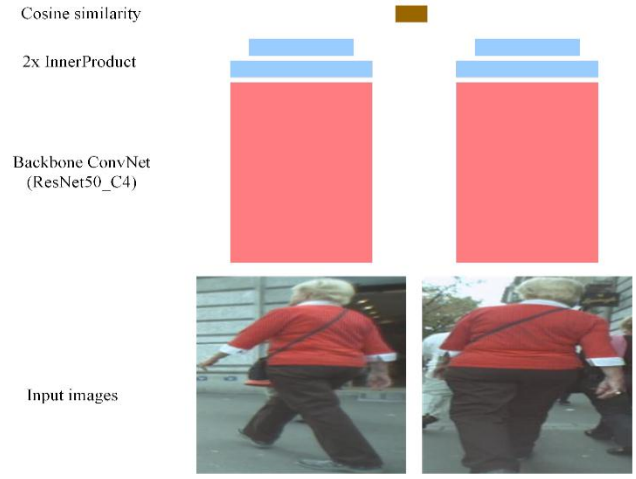


Fig. 4. Illustration of deep Siamese network [8]. Input images are resized to 224 by 224. The two branches are identical until a cosine similarity is calculated between their final feature vectors.

V. CONCLUSION

In this paper, we deal with the failures of tracking guidance models by proposing an online model coupling mechanism. The working status of appearance/motion models is checked online during tracking. Adjustments will be made under the circumstances of the temporary model dysfunction. Motion models are designed to take over when appearance models have low credibility in terms of the association prediction, with the help of a variable coupling coefficient. As for the model update, outliers are prevented from dominating the appearance models with the help of a weighted observation combination. Experiments were carried out on the MOT Challenge benchmark, and the results showed the benefits of our proposed strategies.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, pp. 1097-1105, 2012.

[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, pp. 91-99, 2015.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778, 2016.

[4] B. Babenko, M. H. Yang, and S. Belongie, "Visual tracking with online multiple instance learning," *Computer Vision and Pattern Recognition,* pp. 983-990, 2009.

[5] C. Kim, F. Li, A. Ciptadi, and J. M. Rehg, «Multiple hypothesis tracking revisited," in *Proc. IEEE International Conference on Computer Vision*, pp. 4696-4704, 2015.

[6] S. Tang, M. Andriluka, B. Andres, and B. Schiele, "Multiple people tracking by lifted multicut and person reidentification," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3539-3548, 2017.

[7] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, "MOT16: A benchmark for multi-object tracking," 2016.

[8] B. Cuan, K. Idrissi, and C. Garcia, "Deep siamese network for multiple object tracking," in *Proc. IEEE 20th International Workshop on Multimedia Signal Processing*.

[9] H. V. Nguyen and L. Bai, "Cosine similarity metric learning for face verification," in *Proc. Asian Conference on Computer Vision*, pp. 709-720, Springer, Berlin, Heidelberg, 2010.

Bonan Cuan is a Ph.D. student of the Institut National de Sciences Appliquées in Lyon, France (INSA Lyon) from 2015. He received his B.S. and M.S. degree in electronic and information engineering from Xi'an Jiaotong University, China in 2012 and 2015, respectively. He received his engineer's degree in 2015 from the Université de Technologie de Troyes, France.

Khalid Idrissi received his B.S. and M.S. in 1984 in electrical engineering from INSA Lyon, France. From 1985 to 1991, he worked as an engineer, then as the head of a project in industry. He received his "Agrégation" in electrical engineering in 1994 and has been "professeur agrégé" until 2003 at the French Guyana University then at INSA in Lyon. He received his Ph.D. in 2003 and his "HDR" in 2011 and is now working as an associate professor at the Telecommunication Department of INSA-Lyon since 2004. He leads his research activities in the IMAGINE team of the LIRIS research laboratory. He is mainly working on image analysis and segmentation for image compression, image retrieval, shape detection and identification, and facial analysis.

Christophe Garcia is a full Professor at the Institut National de Sciences Appliquées at Lyon, France (INSA Lyon), teaching computer science in the first cycle department (computer systems, databases, and algorithms); machine learning and pattern recognition in the IT departments. Since July 2015, he has been the Deputy Director of the LIRIS Laboratory (Laboratoire d'InfoRmatique en Image et Systèmes d'information), a joint lab UMR 5205 CNRS/INSA de Lyon/Université Claude Bernard Lyon 1/Université Lumière Lyon 2/École Centrale de Lyon. Since September 2016, he has been the Vice-President for Research in the Information and Digital Society of INSA - Lyon.