JOINT TRACKING AND GAIT RECOGNITION OF MULTIPLE PEOPLE IN VIDEO

Maryam Babaee[⋆] *Gerhard Rigoll*[⋆] *Mohammadreza Babaee*[⋆]

[⋆] Institute for Human-Machine Communication, TU Munich, Germany

ABSTRACT

We propose a novel approach to address the problem of jointly tracking and gait recognition of multiple people in a video sequence. The most state of the art algorithms for gait recognition consider the cases where there is only one person without any occlusion in a very constrained environment. However, in real scenarios such as in airports, train stations, etc, there are many people in the environment that make these algorithms inapplicable. Although first tracking of each person and then gait recognition could be a solution, we argue that the multi-people tracking and the gait recognition in a video are two sub-problems that can help each other. Hence, we propose a joint tracking and gait recognition of multiple people as one framework that can improve gait recognition accuracy and decrease the ID switching in tracking. Experimental results confirm the validity of proposed approach.

Index Terms— Gait recognition, Multi-people tracking

1. INTRODUCTION

As one of the biometric features for human recognition, Gait (the way of natural walking) [1] has drawn significant attention in recent years, since it can be used to recognize people from large distances, while other biometric features such as face, fingerprint, and iris [2, 3, 4] are not available. Here, a sequence of images showing a person walking is analyzed as input data. This is the assumption of many proposed gait recognition methods during the past years. However, they are incapable when there are several people walking/running in a uncontrolled environment. Additionally, in multi-people gait recognition, we do not know the matching between targets in two consequent frames in order to make a gait cycle. Thus, first we need to determine the corresponding targets along the frames which can be regarded as an individual (i.e., solving tracking problem). Consequently, multi-people gait recognition and the multi-people tracking problems are closely related.

In this paper, we propose a method that combines multi-people tracking and gait recognition in one framework. Basically, both problems have been treated as independent problems. Here, we claim that both problems are closely related. To this end, we propose two lines of process, one for tracking and the other for gait recognition, that run in parallel and the output of each other helps the other.

The rest of the paper is organized as follows. In Section 2, we provide an overview of related works. Section 3 explains our proposed approach and the detail of the joint formulation of Gait recognition an tracking of multiple people. In Section 4, we show the efficiency of the algorithm by conducting several experiments on real datasets. Finally, in Section 5, we provide a short summary and draw our conclusion.

2. RELATED WORK

The gait recognition approaches can be divided into two groups; 1) the model-based and 2) the appearance-based techniques. In the first group, a 3D pre-defined model is optimized, while in the second group, appearance features are extracted from one single image. Since finding a 3D model requires high resolution images, we concentrate on the second type approaches. In this category, different approaches have been proposed for gait recognition including applying Hidden Markov Models (HMM) on an image silhouette for classification [5]. As a simple feature, the average over a complete gait cycle silhouettes, the so-called Gradient Energy Image (GEI) is used in [6]. To enhance the feature, an extended version of GEI has been proposed in [4], where alpha-matte as a foreground subtraction algorithm is used before computing gradient histogram of energy image. Most approaches proposed for gait recognition are able to recognize a person from images containing only one person. To our knowledge, Gait Recognition for multiple people has not been investigated combined with tracking. All widely used standard datasets particularly CMU Motion of Body (MoBo) [7], the USF Gait Based Human ID Challenge [8] and Casia-B [9] contain only one person walking in a constrained environment. To identify a person in a multiple people video sequence, obviously, it is required to first track the underlined person along the video. Most proposed methods for multiple people tracking are based on tracking by detection paradigm [10, 11]. Here, people in one frame are detected by a pre-trained human detector and then are tracked by a data association method which can be formulated as a network flow problem [11]. Among many algorithms for solving the max-flow min-cut problem, we use the Binary Integer Programming for data association.

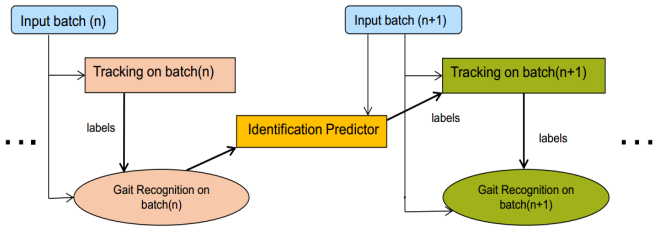


Fig. 1. The flowchart of the proposed joint tracking and gait recognition

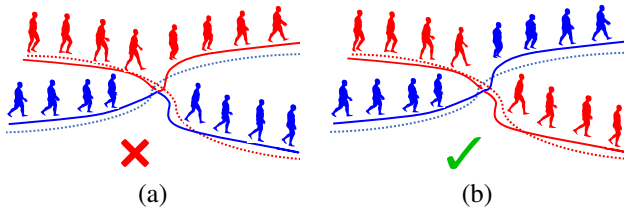


Fig. 2. The problem of ID-switching in tracking. (a) ID-switching happens and the gaits are not consistent. (b) there is no ID switching and the gaits are consistent. The dot lines show ground truth trajectories and the solid lines show the trajectories obtained by tracking.

3. APPROACH

In both tracking and Gait recognition problems, the goal is to assign a unique ID to each person detected in a video sequence. In our proposed approach, we combine these two problems into one single framework. As depicted in Fig. 1, the proposed framework contains two processing lines. The first line is for tracking and the second line for Gait recognition. We argue that these two problems can help each other to improve the accuracy of ID assignment. Here, first the tracking algorithm is run on the first batch of the data (about first 50 frames of the data sequence) and then each detection gets an ID number. This information is used to train a SVM (second processing line). Before running the tracking algorithm on the second batch, the trained SVM predicts an ID to each detection of this batch as primary ID. These primary ID numbers are used in the network flow of tracking algorithm applied to second batch. This prior gait information helps reducing ID-switch error (See Fig. 2). The detailed information about tracking and gait recognition have been explained in Sections 3.1 and 3.2, respectively.

3.1. Tracking multiple people

In tracking, detection boxes are used as our main input data. A 2D detection is defined by $\boldsymbol{D}_i = (x_i, s_i, t_i)$, where x_i is the position, s_i the size of the detection, and t_i the time. The set of all detections is denoted as D. A track is defined as $\mathcal{T}_u = \{\mathcal{D}_{u_1}, \mathcal{D}_{u_2}, ..., \mathcal{D}_{u_n}\}$. The tracking problem is achieved by finding an optimal set of tracks $\mathcal{T}^*$, which has the Maximum a-Posteriori (MAP) probability given detections ($\mathcal{D}$) and gait cycles ($\mathcal{G}$) as observations. Since detections and gait cycle are conditionaly independent, the MAP formulation would be;

$$\begin{aligned}\mathcal{T}^* &= \underset{\mathcal{T}}{\operatorname{argmax}}\, P(\mathcal{T}|\mathcal{D},\mathcal{G}) \\ &= \underset{\mathcal{T}}{\operatorname{argmax}}\, P(\mathcal{D},\mathcal{G}|\mathcal{T})P(\mathcal{T}) \\ &= \underset{\mathcal{T}}{\operatorname{argmax}}\, P(\mathcal{D}|\mathcal{T})P(\mathcal{G}|\mathcal{T})P(\mathcal{T})\end{aligned} \tag{1}$$

The posterior probability can be written as the multiplication of the likelihood and prior probabilities according to the Bayes rule. We further assume non-overlapping trajectories, i.e., one detection $\mathcal{D}_k$ can only be part of at most one trajectory:

$$\mathcal{T}_u \cap \mathcal{T}_v = \emptyset,\ \forall u \neq v \tag{2}$$

and two gait cycles must not intersect

$$\mathcal{G}_k \cap \mathcal{G}_l = \emptyset,\ \forall k \neq l,\ \forall \mathcal{G}_k, \mathcal{G}_l \in \mathcal{T}. \tag{3}$$

With this non-overlapping assumptions, the individual likelihood probabilities $P(\mathcal{D}_k|\mathcal{T})$ are conditional independent and the individual prior probabilities $P(\mathcal{T}_u)$ are independent as well. So the MAP formulation can be factorized as:

$$\begin{aligned}\mathcal{T}^* = \underset{\mathcal{T}}{\operatorname{argmax}} \prod_{\mathcal{D}_k \in \mathcal{D}} P(\mathcal{D}_k|\mathcal{T}) \prod_{\mathcal{G}_i \in \mathcal{G}} P(\mathcal{G}_i|\mathcal{T}) \prod_{\mathcal{T}_u \in \mathcal{T}} P(\mathcal{T}_u) \\ s.t. \mathcal{T}_u \cap \mathcal{T}_v = \emptyset,\ \forall u \neq v\end{aligned} \tag{4}$$

The terms in (4) are defined as follows:

$$\begin{aligned}P(\mathcal{T}_u) &= P(\{(\mathcal{D}_{u_0}, \mathcal{D}_{u_1}, ..., \mathcal{D}_{u_{n_u}}\}) \\ &= P_{en}(\mathcal{D}_{u_0}) P_{link}(\mathcal{D}_{u_1}|\mathcal{D}_{u_0}) ... P_{ex}(\mathcal{D}_{u_{n_u}})\end{aligned} \tag{5}$$

The MAP problem can be reformulated into a cost-flow graph and solved exactly in an Binary Integer Programming (BIP) [12, 13]:

$$\begin{aligned}\mathcal{K}^* = \underset{\mathcal{K}}{\operatorname{argmin}} &\sum_{k \in \mathcal{K}} C_k f_k + \sum_{i \in \mathcal{I}} C_{\mathcal{G}_i} \mathcal{G}_i + \sum_{k \in \mathcal{K}} C_{en,k} f_{en,k} \\ &+ \sum_{k \in \mathcal{K}} C_{k,l} f_{k,l} + \sum_{k \in \mathcal{K}} C_{k,ex} f_{k,ex} \\ s.t. \quad f_{en,k} + \sum_l f_{l,k} &= f_k = f_{ex,k} + \sum_l f_{k,l},\ \forall k\end{aligned} \tag{6}$$

Here, f_k denotes whether the detection is chosen to be a part of the track (1) or not (0). To obtain the optimal subset of the tracks, the non-overlap constraint between tracks is also used here. This means, two different tracks cannot contain the same detection.

The costs C are merged from the MAP formulation. We use the entrance probability $C_{en,k}$ and exit probability $C_{k,ex}$ as they are defined in [12]. C_k is the detection cost which takes the influence of false detection alarm into account:

$$C_k = \omega_{det}\Big(\log\frac{\beta}{1-\beta}\Big) \tag{7}$$

where β is the false positive rate of the detector and ω_{det} is the weight factor of detection cost. Each $\mathcal{G}_i$ is a set of detections shows a gait cycle of an individual and it's cost is defined based on the predicted score of $\mathcal{G}_i$ obtained by the SVM.

$$C_{\mathcal{G}_i} = -\log(Score_{SVM}(\mathcal{G}_i)) \tag{8}$$

The cost of a transition edge is

$$C_{k,l} = \omega_{geo}Z_{geo}(k,l) + \omega_{gait}Z_{gait}(k,l) \tag{9}$$

which consists of geometric distance between two detections (Z_{geo}) as well as the consistency of their gait features (Z_{gait}).

$$\begin{aligned} Z_{gait}(k,l) &= -\log(P_{trans-gait}(D_l|D_k)) \\ &= -\log(1 - dist(D_l,D_k)/dist_{max}) \end{aligned} \tag{10}$$

Here, $P_{trans-gait}$ is the transition probability based on gait feature consistency. The dissimilarity of the gait features is defined as

$$dist(D_l,D_k) = \begin{cases} 0, & \text{if } gaitID_l = gaitID_k \\ 3, & \text{if } gaitID_l \neq gaitID_k \\ 1, & \text{if no ID has been assigned} \end{cases} \tag{11}$$

The cost value Z_{geo} is computed by

$$Z_{geo}(k,l) = -\log(P_{trans-geo}(D_l|D_k,\Delta t)P(\Delta t)) \tag{12}$$

where, $P_{trans-geo}$ is the transition probability based on geometric distance defined by

$$P_{trans-geo}(D_l|D_k,\Delta t) = \mathcal{F}(||D_k - D_l||, 0, \frac{v_{max}}{f}\Delta t) \tag{13}$$

where Δt and v_{max} are the frame gap and maximum velocity of walking, respectively. $\mathcal{F}$ is a decreasing function to map the distances to the costs which is defined as:

$$\mathcal{F}(d, d_{min}, d_{max}) = \frac{1}{2}\text{erfc}(4\frac{d - d_{min}}{d_{max} - d_{min}} - 2) \tag{14}$$

that maps the distance to the range $[0,1]$. According to [14], the temporal frame gap probability $P(\Delta t)$ is defined as:

$$P(\Delta t) = \begin{cases} \gamma^{(\Delta t - 1)}, & \text{if } 1 \leqslant \Delta t \leqslant \Delta t_{max} \\ 0, & \text{otherwise} \end{cases} \tag{15}$$

where the maximal frame gap of two detections is denoted as Δt_{max} and γ is the false negative rate of the detector.

3.2. Gait Recognition

To improve the tracking algorithm, we utilize the gait features to improve the association of the detections of the same person along the sequence. After applying tracking on batch n, the detections with assigned labels are given to a Support Vector Machine (SVM) as training data. In training, first the Gradient Histogram Energy Image (GHEI)[4] of a full gait cycle for each person is extracted and fed into the SVM. The GHEI uses the gradient histograms at all locations of the original image. Therefore, the edge information inside the boundary of the objects are used. The background information of the HOG images leads to a degraded recognition performance. To overcome this issue, we use alpha matte-GHEI feature vector proposed by [4]. The SVM gets alpha-GHEI images of detection boxes labeled by the tracking algorithm as training data. After training, the SVM predicts the label of all possible sub-trajectories for the next batch (i.e. Batch $n+1$).

3.2.1. Generating Gait Hypotheses

As the number of detections in each frame increases, the number of all possible trajectories will increase exponentially. Since there will be many hypotheses coming from different combinations of detections along a gait cycle, we only take those valid trajectories which are more likely to belong to a person. Therefore, in a trajectory, each detection should be close enough to its consecutive detections. Meanwhile, the sub-trajectories must not have overlaps. By considering these constraints, the number of trajectories of test data will drastically decrease and therefore the SVM predicts faster. Next we assume that two different gait cycle $\mathcal{G}_k$ and $\mathcal{G}_l$ cannot contain one or more same detections:

$$\mathcal{G}_k \cap \mathcal{G}_l = \emptyset,\ \forall k \neq l,\ \forall \mathcal{G}_k, \mathcal{G}_l \in \mathcal{T} \tag{16}$$

3.2.2. Gait Prediction

After selecting the more probable sub-trajectories, we identify their ID using the SVM classifier trained by the tracking result of previous batches as training data. In each cycle, there has to be at most one sub-trajectory for each person which is represented as a specific ID number. The SVM gives a score s_i to each sub-trajectory showing how likely this sub-trajectory belongs to the class i. If two different sub-trajectories have the same ID, one that has higher score is selected as the trajectory corresponding to that person.

4. EXPERIMENTS

To evaluate the performance of the proposed approach, we conducted experiments on two widely used datasets; 1) PETS2009-S2L1 and 2) TUD-Stadtmitte. The first one contains 795 frames with the frame rate of 7, and the second one

is consisted of 179 frames with the frame rate of 25.

Evaluation Metrics: Since we do both tracking and gait recognition, we used evaluation metrics for both problems. We report the frequently used CLEAR MOT metrics in [15], including Multiple Object Tracking Accuracy (MOTA), Multiple Object Tracking Precision (MOTP), Mostly Tracked trajectories(MT, $> 80\%$ overlap), Mostly Lost (ML, $< 20\%$ overlap) and Partly Tracked (PT) and ID-switch for tracking problem. For Gait recognition, we report the accuracy of gait recognition.

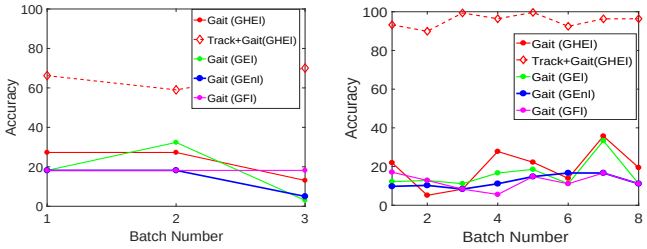
Settings: The initial target-ID by gait plays an important role in the data association, so we set $\omega_{gait} = 0.5$. The false positive rate β is related to the detection quality, and we set $\beta = 0.05$. Since the frame rates for the two data sets PETS-S2L1 and TUD-Stadtmitte are different, we set the gait cycle γ differently for each one. The gait cycle for the PETS-S2L1 dataset is $\gamma = 15$, and for the TUD-Stadtmitte is $\gamma = 25$.

To evaluate the pure gait recognition (i.e., without tracking), we used the groundtruth IDs to build true trajectories. After training the SVM from the previous batch, the classifier predicts the ID of each detection for the next batch. We applied different gait feature extraction methods for gait recognition including GEI [16], GHEI, GFI (Gait Flow Image), GEnI (Gait Entropy Image) for comparison.

Results and Discussion: We report the accuracy of gait recognition with and without tracking for different gait features presented in Fig. 3. Evidently, the joint method outperforms the pure gait methods, while gait recognition algorithms alone do not perform well on video surveillance datasets where the people are observed from a long distance. However, on these datasets, tracking algorithm can help the gait recognition. Fig. 4 represents two samples of the result of segmentation (used for computing GHEI) and tracking. Additionally, quantitative tracking results compared with other methods for S2.L1 and TUD-Stadtmitte shown in Table 1 confirm that we achieve less ID switching while keeping comparable MOTA and MOTP. Here, we report the tracking results with gait (Ours-Track-Gait) and without gait (Ours-Track). The results show that Gait feature helps tracking achieve less ID switching.

5. CONCLUSION

We have addressed the problem of joint tracking and gait recognition of multiple people in video sequences. We argued that both tracking and gait recognition are closely related to each other and could help each other if we can have both subproblems in one framework. To this end, a network flow was proposed that involves gait features to enhance the performance of trajectory extraction and person identification. Our experiments conducted on two datasets confirm that gait feature could decrease the ID switching of the tracking problem and tracking can find the trajectory of people. As future work, one might utilize other appearance based features such as human pose to improve the performance.



(a) TUD-Stadtmitte (b) S2L1-View 1

Fig. 3. The comparison of the two types of methods; 1) pure gait recognition (Gait) with different gait features and 2) joint gait-tracking (Track + Gait).

(a)TUD-Stadtmitte (b)PETS-S2L1 (View 1)

Fig. 4. Identification results on the two datasets; The input images(first row), the segmented images [17] (second row), and the recognition results (third row).

Seq.	Method	MOTA	MOTP	IDS	MT	PT	ML
S2.L1	Berclaz et al.[18]	80.3	72.0	13	73.9	17.4	8.7
	Milan et al. [19]	90.3	74.3	22	78.3	21.7	0
	Pirsiavash et al. [11]	77.4	74.3	57	60.9	34.7	4.3
	Andriyenko et al. [20]	88.3	79.6	18	82.6	17.4	0
	Chari et al. [21]	85.5	76.2	56	94.7	0	0
	Milan et al. [22]	85.3	77.5	9	100	0	0
	Ours-Track	96.0	81.3	11	94.74	5.26	0
	Ours-Track-Gait	**96.0**	**81.8**	**8**	**100**	**0**	**0**
Stadtmitte	Milan et al. [19]	56.2	61.6	15	44.4	55.5	0
	Chari et al. [21]	51.6	**61.7**	15	20.0	**80.0**	0
	Ours-Track	51.21	60.58	13	30.0	60.0	10.0
	Ours-Track-Gait	**63.06**	57.77	**9**	**50.0**	50.0	**0**

Table 1. Tracking results of S2L1 and TUD-Stadtmitte.

6. REFERENCES

[1] J. E. Boyd and J. Little, "Biometric gait recognition," in *Advanced Studies in Biometrics*, pp. 19–42. Springer, 2005.

[2] X. Xing, K. Wang, T. Yan, and Z. Lv, "Complete canonical correlation analysis with application to multi-view gait recognition," *Pattern Recognition*, vol. 50, pp. 107–117, 2016.

[3] T. Wolf, M. Babaee, and G. Rigoll, "Multi-view gait recognition using 3d convolutional neural networks," in *Image Processing (ICIP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 4165–4169.

[4] M. Hofmann and G. Rigoll, "Exploiting gradient histograms for gait-based person identification," in *Image Processing (ICIP), 2013 20th IEEE International Conference on*. IEEE, 2013, pp. 4171–4175.

[5] A. Kale, A. Sundaresan, A.N. Rajagopalan, N.P. Cuntoor, A. K. Roy-Chowdhury, V. Kruger, and R. Chellappa, "Identification of humans using gait," *IEEE Transactions on image processing*, vol. 13, no. 9, pp. 1163–1173, 2004.

[6] J. Man and B. Bhanu, "Individual recognition using gait energy image," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 2, pp. 316–322, 2006.

[7] R.h Gross and Jianbo Shi, "The cmu motion of body (mobo) database," 2001.

[8] S. Sarkar, P. Jonathon Phillips, Z. Liu, I. Robledo Vega, P. Grother, and K.W. Bowyer, "The humanid gait challenge problem: Data sets, performance, and analysis," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 2, pp. 162–177, 2005.

[9] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*. IEEE, 2006, vol. 4, pp. 441–444.

[10] A.R. Zamir, A. Dehghan, and M. Shah, "Gmcp-tracker: Global multi-object tracking using generalized minimum clique graphs," in *Computer Vision–ECCV 2012*, pp. 343–356. Springer, 2012.

[11] H. Pirsiavash, D. Ramanan, and C. Fowlkes, "Globally-optimal greedy algorithms for tracking a variable number of objects," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 1201–1208.

[12] M. Hofmann, D. Wolf, and G. Rigoll, "Hypergraphs for joint multi-view reconstruction and multi-object tracking," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE International Conference on*, pp. 3650–3657.

[13] T. Kroeger, R. Dragon, and L. Van Gool, "Multi-view tracking of multiple targets with dynamic cameras," in *Pattern Recognition*, pp. 653–665. Springer, 2014.

[14] Li Zhang, Yuan Li, and Ramakant Nevatia, "Global data association for multi-object tracking using network flows," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[15] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The clear mot metrics," *Image and Video Processing, EURASIP Journal on*, vol. 2008, pp. 1–10, 2008.

[16] Ju Man and Bir Bhanu, "Individual recognition using gait energy image," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 2, pp. 316–322, 2006.

[17] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr, "Conditional random fields as recurrent neural networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1529–1537.

[18] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua, "Multiple object tracking using k-shortest paths optimization," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 9, pp. 1806–1819, 2011.

[19] A. Milan, K. Schindler, and S. Roth, "Detection-and trajectory-level exclusion in multiple object tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3682–3689.

[20] A. Andriyenko, K. Schindler, and S. Roth, "Discrete-continuous optimization for multi-target tracking," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 1926–1933.

[21] V. Chari, S. Lacoste-Julien, I. Laptev, and J. Sivic, "On pairwise costs for network flow multi-object tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5537–5545.

[22] A. Milan, L. Leal-Taixé, K. Schindler, and I. Reid, "Joint tracking and segmentation of multiple targets," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5397–5406.