Multi-class Classification and Clustering based Multi-object Tracking

Nii Longdon Sowah, Qingbo Wu, Fanman Meng

***Abstract*— In this paper, we study the challenging problem of tracking multiple moving objects by a single camera. In our approach, a novel and efficient way to obtain an appearance and temporal based tracklet affinity model is proposed. We also propose to formulate the tracking problem as a classification task, where we classify tracklets into multi-classes, jointly across space and time. In our framework, we formulate our objective as a decision function with a Bayesian classifier constraint. We learn a model which seeks to maximize the decision function whilst minimizing our constraint. We estimate the probability density function of each class using a Gaussian probability density function, which reduces our misclassification loss. Our tracklet generation method minimizes the effect of missed detections and false positives. The proposed method emphasizes the effectiveness of multi-class SVMs (Support Vector Machines) in Multi-object tracking (MOT). Experimental results on three widely used Multi-Object Tracking datasets show that our method outperforms several state-of-the-art approaches in multi-object tracking.**

***Index Terms*—classifier; multi-class; support vector machine; tracklet.**

I. Introduction

Tracking multiple objects in videos is an important problem in computer vision due to its wide applications in various video analysis scenarios, such as surveillance, sports analysis, robot navigation and autonomous driving. Recent progress on Multi-Object Tracking (MOT) has focused on the tracking-by-detection strategy, where the object detections from an object detector are linked to form trajectories of the targets. Various machine learning algorithms have been explored to track multiple objects [1] – [3], [20]–[22], which can be categorized into offline-learning methods and online-learning methods. Online learning conducts learning during tracking. A common strategy is to construct positive and negative training samples, and then to train a similarity function for data association [1]–[2]. Such methods update their training parameters during tracking. Online-learning is able to utilize features based on the status and the history of the target.

However, such methods are sensitive to incorrect training examples produced in the tracking results from previous frames, and these errors can be accumulated and result in tracking drift. Offline learning methods learn model parameters before tracking. These methods generate tracklets from detections and perform data association over the tracklets [19]–[20]. Several offline-learning MOT methods have been proposed which typically construct a target-specific classifier to associate the detections with specific targets such as Shu et al. [3] and Breitenstein et al. [4]. The method proposed by Breitenstein et al. [4] proposes a particle-based framework in which detections and intermediate detections' confidences are used to propagate the particles. Additionally, they employ a target-specific classifier to associate the detections with the trackers. Their method is however, not robust to false positives and their use of detection confidence for data association is unreliable. Shu et al. [3] use an extended part-based human detector on every frame to extract the part features from all detections. They then train person-specific SVM classifiers using the detections, and consequently classify the new detections. This approach is computationally more expensive and combining multiple binary classifier scores increases the complexity of the classification task in solving the data association problem. Wang et al[17]–[18] proposed two approaches for learning parameters of min-cost flow MOT using quadratic trajectory interactions including suppression of overlapping tracks and contextual cues about co-occurrence of different objects. In [17] they utilized structured prediction with a tracking-specific loss function to learn the complete set of model parameters. They found an optimal set of tracks under a quadratic model objective based on an LP relaxation and a greedy extension to dynamic programming that handles pairwise interactions. In [18] they augmented a standard min-cost flow objective with quadratic terms between detection variables to model pairwise interactions between different tracks. These approaches however, differ from our proposed method in that, they did not include appearance features and also used a structured SVM compared to our method which uses a kernelized SVM with appearance features.

In this paper, we propose a novel MOT algorithm which formulates the tracking task as classifying detections and clustering similar detections, as described in Fig. 1. We also propose a multi-class SVM model which is able to classify multiple objects with high accuracy and thereby, reduce the computational complexity of using target specific classifiers

Manuscript received March 19, 2017; revised March 30, 2017. This work was supported in part by National Natural Science Foundation of China under grant number 61601102.

Nii Longdon Sowah is with the Department of Electronic Engineering, University of Electronic Science and Technology of China ,611731, Chengdu, Sichuan Province, China (phone: +86-18684027474 ; email : longdon001@yahoo.co.uk).

Qingbo Wu is with the Department of Electronic Engineering, University of Electronic Science and Technology of China,611731, Chengdu, Sichuan Province, China (e-mail: wqb.uestc@gmail.com).

Fanman Meng is with the Department of Electronic Engineering, University of Electronic Science and Technology of China,611731, Chengdu, Sichuan Province, China (e-mail: fmmeng@uestc.edu.cn)

ISBN: 978-988-14047-4-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

in MOT. The generated classes are clustered and interpolated to obtain the final target trajectories. The contribution of this work is three fold: firstly, we propose a multiclass classification model to solve the MOT task. Secondly, we develop a new trajectory generation method by clustering the classified objects. Thirdly we prove by extensive experiments on three datasets that our method outperforms several state-of-the-art MOT algorithms.

II. Proposed Framework

Our method adopts a two layer tracking framework, in which initial shorter tracklets are merged into final trajectories. Based on detection results obtained from a reliable object detector, we generate a classification model via a couple of steps as shown in Fig. 1. In the first step, 3D color histogram and LBP features for each object detection are extracted. The color histogram features help distinguish objects of different colors, whilst the LBP features help to distinguish objects with different texture. These features are used to train a linear SVM model. In the next step, we divide the video into 10 frames each and merge the detections in each segment. According to our merging criteria, we merge consecutive detections that have an overlap ratio of 60% between frames. Tracklets that are shorter than 5 detections within a segment are discarded to account for false positives or missed detections.

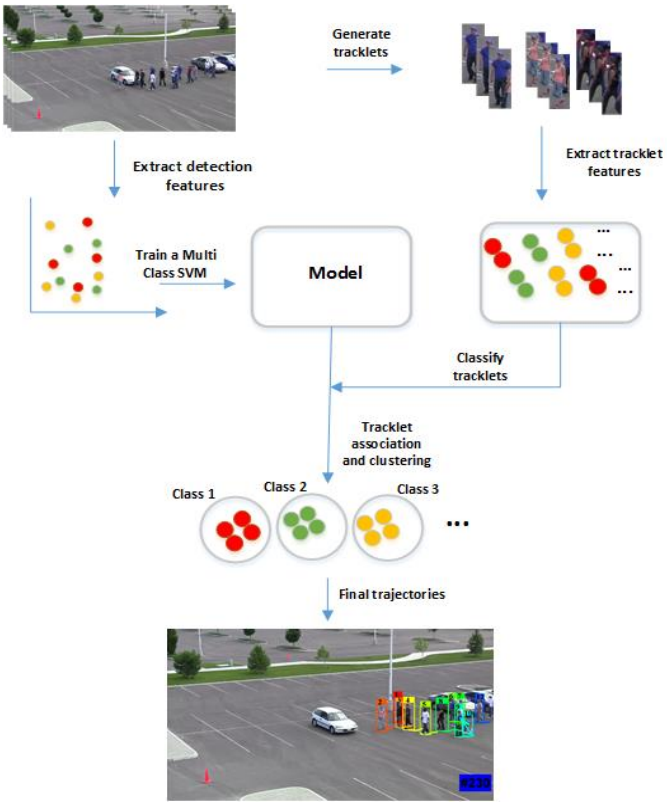

Fig. 1. Our proposed framework

Due to the fact that there is uniform motion of the pedestrians, we can use the mean detection coordinates of a tracklet to represent it. We classify our tracklets using the model generated by training the detections. The negative samples used in the classification are the background detections. This helps to improve our multi-class classification score. We consider each class as a cluster and extrapolate the mean detections of the tracklets to form the final trajectories.

A. Tracklet Generation

Tracklets are the input to our tracking algorithm. Tracklets are used based on the fact that in most tracking scenarios, there's not much change in object articulation between successive frames. It is worth to mentioning that tracklets have been previously used as reliable inputs in many tracking algorithm [5]–[7]. In our method, tracklets help reduce the computational complexity. These tracklets are found using an overlap criteria where bounding boxes that overlap more than 60% in consecutive frames are connected. We use the general polygon clipping algorithm by Vatti [8] to compare detections and merge those that meet our overlap criteria. For every tracklet, we obtain its bounding box coordinates, its frame numbers, and its label. We divide the video into segments of 10 frames each. We generate tracklets for each segment.

B. Person Classification

We mentioned in the introduction that one advantage of the proposed method is the fact that we classify all objects simultaneously with a multi-class SVM. To train the model of our SVM, we extract features from every detection in each frame and concatenate them as a feature vector. We chose Local Binary Pattern (LBP) and color histogram features since they are highly discriminative of texture and color. Since we use a multi-class SVM, each object detection is a positive sample and we have different classes for positive samples. We do not use the detections of other objects as negative samples. To generate negative samples for our classifier, we augment our positive samples with background detections to account for negative samples. This helps improve the classifier's discrimination to the background.

III. Tracklet Association

Given a tracklet v, the probability that it belongs class w_j from a total number of W classes is given by the decision function

$$q_j(v) = \max p(v / w_j) P(w_j) \tag{1}$$

$$\text{s.t } r_j(v) = \min \sum_{k=1}^{W} L_{kj}\, p(w_k / v) \tag{2}$$

where the constraint in (2) is the Bayes classifier (minimum misclassification loss function) and L_{kj} is the misclassification cost, which is the cost of classifying tracklet v as belonging to class j instead of class k. From basic probability theory, we can rewrite the misclassification loss function as

$$r_j(v) = \frac{1}{p(v)} \sum_{k=1}^{W} L_{kj}\, p(v / w_k) P(w_k) \tag{3}$$

ISBN: 978-988-14047-4-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

$$= \sum_{k=1}^{W} L_{kj}\, p(v / w_k) P(w_k) \quad (4)$$

For any two classes i and j, assuming the maximum misclassification cost is 1 and the minimum is 0, the misclassification cost can be given by

$$L_{ij} = 1 - \delta_{ij} \quad (5)$$

where $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$ substituting (5) into (3) gives us

$$r_j(v) = \sum_{k=1}^{W} (1 - \delta_{kj})\, p(v / w_k) P(w_k) \quad (6)$$

$$= p(v) - p(v / w_j) P(w_j) \quad (7)$$

which is similar to the decision function in (1). We solve $p(v / w_j)$ using a classifier for Gaussian pattern classes.

The Gaussian density of the vectors in the j^{th} class has the form

$$p(v / w_j) = \frac{1}{(2\pi)^{1/2} |C_j|^{1/2}} e^{-1/2 (v - m_j)^T C_j^{-1} (v - m_j)} \quad (8)$$

with mean m_j and covariance C_j matrix. The mean and covariance matrix can be expressed in the form

$$m_j = E_j\{v\} \quad (9)$$

$$C_j = E_j\{(v - m_j)(v - m_j)^T\} \quad (10)$$

where E_j is the expected value of v in class j. Approximating the expected value E_j, yields an estimate of the mean and covariance matrix.

Given any two tracklets, y_i and y_j, we cluster them by the clustering probability given as

$$P(y_i, y_j) = p(y_i / w_i) \cdot p(y_j / w_i) \cdot A(y_i, y_j) \quad (11)$$

where the first two terms is are the probabilities that y_i and y_j belong to the same class and $A(y_i, y_j)$ is the temporal difference between the tracklets given by

$$A(y_i, y_j) = \begin{cases} 1, if & \Delta t > 0 \\ 0, & otherwise \end{cases} \quad (12)$$

We smooth the tracklet clusters to obtain the final trajectories.

IV. Experiments

In our evaluation, we focused on pedestrian tracking due to its importance and popularity in most MOT algorithms. We extensively experimented on the proposed method using TUD Stadtmitte (Stadtmitte) [11], Parking Lot 1 (PL1) and Parking Lot 2 (PL2) [12] datasets. The experimental datasets provide a wide range of significant challenges including occlusion, objects with same color and cluttered background. We used LIBSVM [23] for our multi-class classification. In all the sequences, we only use the visual information and do not use any scene knowledge such as the camera calibration. We used the groundtruth detections for training the classifier. Also tracking is performed on the entire viewing range of the camera. We compared our method with the state-of-art trackers, borrowing the numbers from the authors' papers. We adopt the commonly used CLEAR MOT metrics [21] which is the standard in comparing MOT algorithms:

- Mostly tracked trajectories (MT), the percentage of trajectories that are successfully tracked for more than 80% divided by ground truth.
- ID switches (IDS): The total number of times that a tracked trajectory changes its matched GT identity.
- Recall (Rec.): The number of correctly matched detections divided by the total number of detections in ground truth.
- Precision (Prec.): The number of correctly matched detections divided by the number of output detections.
- Multi-Object Tracking Accuracy (MOTA): A measure of tracking accuracy that takes into consideration, false positive, false negatives and ID switches.
- Multi-Object Tracking Precision(MOTP): The average bounding box overlap over all tracked targets as a measure of localization accuracy

V. Discussion

We present our tracklet classification accuracy in Table 1. Because our proposed method classifies all the tracklets simultaneously, we have varied accuracy depending on the dataset. Stadtmitte obtains the highest accuracy of 88%, which we attribute to the lower viewing angle of the objects in the video. The camera is quite close to the objects resulting in bigger bounding box sizes and more detailed features.

Tables 2, 3, and 4 show the tracking results on Parking Lot 1, Parking Lot 2 and Stadtmitte respectively. Our proposed method achieves superior results compared to the state-of-the-art on MOTA and MOTP. However, for the Parking Lot 1, Parking Lot 2 datasets, our method tracked less trajectories completely, which could be attributed to our tracklet generation scheme. However, it doesn't affect the final accuracy as, our method obtained less ID switches.

We show the qualitative results of our method in Fig. 2.

TABLE I
TRACKLET CLASSIFICATION ACCURACY

Dataset	Classification accuracy	Number of classes
PL1	75.3%	15
PL2	61%	13
Stadtmitte	88%	11

ISBN: 978-988-14047-4-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

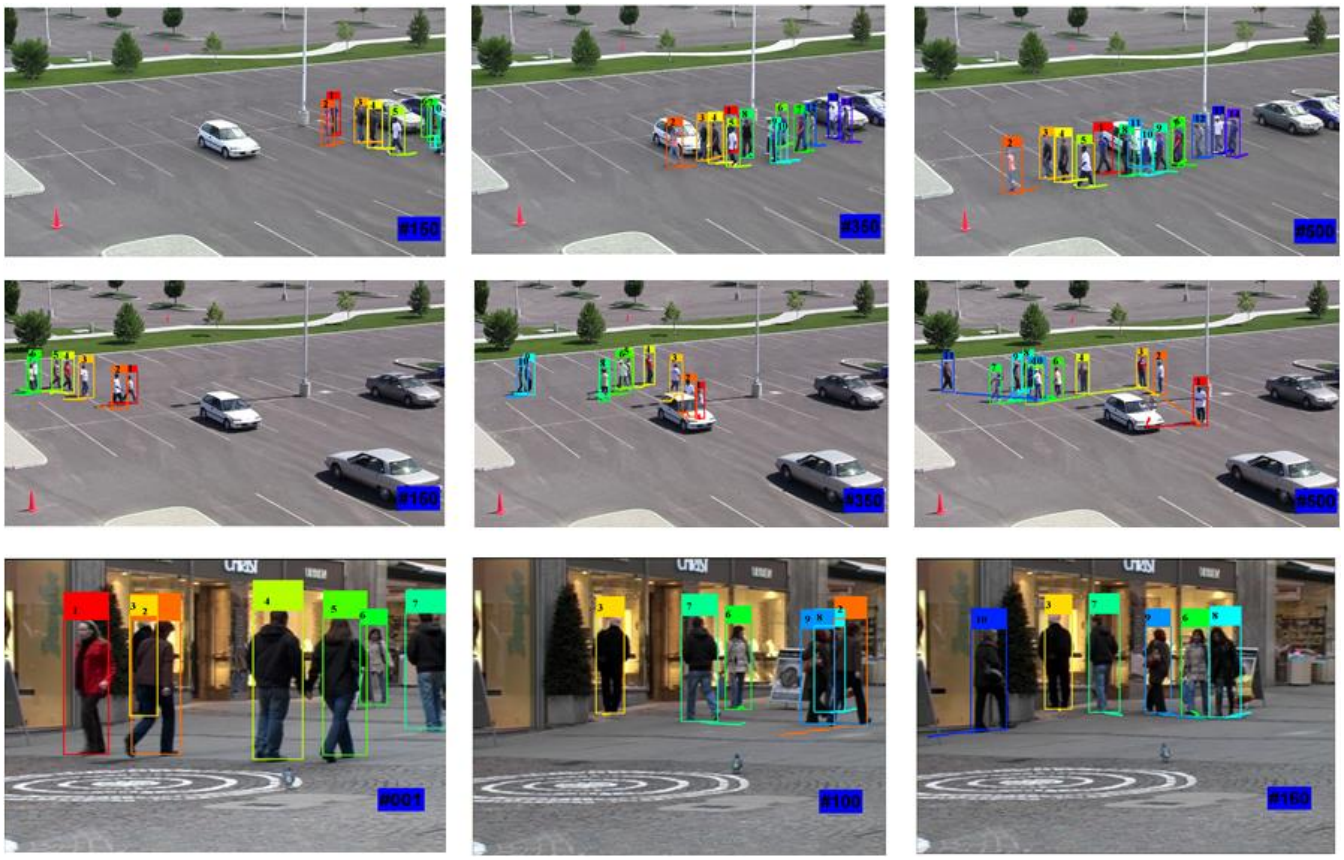

Fig. 2. Tracking results of our tracking approach. Top row shows the Parking Lot 1 dataset, middle row shows the Parking Lot 2 dataset and, the bottom row shows the Stadtmitte dataset

TABLE II
TRACKING RESULTS ON THE PL1 DATASET

Method	Rec.	Prec.	IDS	MT	MOTA	MOTP
PBPO[3]	-	-	-	-	74.1	79.3
GMMCP[10]	-	-	4	13	92.9	73.6
H2T[12]	-	-	21	11	88	81.9
SD[13]	96.1	95.4	18	13	91.4	77.4
Ours	97.6	98.1	3	9	95.7	91.8

TABLE III
TRACKING RESULTS ON THE PL2 DATASET

Method	Rec.	Prec.	IDS	MT	MOTA	MOTP
CMOT[2]	-	-	61	10	80.7	58
TINF[14]	-	-	0	12	89.3	66.3
GMMCP[10]	-	-	7	11	87.6	58.1
Ours	97.0	97.7	15	9	94.5	83.6

TABLE IV
TRACKING RESULTS ON THE STADTMITTE DATASET

Method	Rec.	Prec.	IDS	MT	MOTA	MOTP
DTLE[15]	69.1	85.6	15	4	56.2	61.6
PCNF[16]	59.6	89.9	15	2	51.6	61.6
GMMCP[10]	-	-	0	8	82.4	73.9
Ours	93.1	95.6	19	8	87.2	97.5

VI Conclusion

In this paper we propose multi-object tracking as a multi-class classification and clustering task. We first train a multi-class classifier from the detections and use the model to predict the appearance and temporal based tracklets. Then we cluster the tracklets according to our clustering approach. The clustered tracklets are then interpolated to form the final trajectories. The experimental results show the proposed method performs outperforms state-of-the-art methods

Acknowledgment

Nii Longdon Sowah thanks Dr. Qingbo Wu and Dr. Fanman Meng of the Intelligent Visual Information Processing and Communication Laboratory of the Department of Electronic Engineering, University of Electronic Science and Technology of China for their guidance and support.

References

[1] S. Kim, S. Kwak, J. Feyereisl, and B. Han. "Online multi-target tracking by large margin structured learning." In *Asian Conference on Computer Vision*, pp. 98-111. Springer Berlin Heidelberg, 2012..

[2] S.-H. Bae and K.-J. Yoon. "Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1218-1225. 2014.

[3] G. Shu, A. Dehghan, O. Oreifej, E. Hand, and M. Shah. "Part-based multiple-person tracking with partial occlusion handling." In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 1815-1821. IEEE, 2012.

[4] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. V. Gool. "Robust tracking-by-detection using a detector confidence particle filter." In *Computer Vision, 2009 IEEE 12th International Conference on*, pp. 1515-1522. IEEE, 2009

[5] S. Chen, A. Fern, and S. Todorovic. "Online multi-person tracking-by-detection from a single, uncalibrated camera." In *CVPR*. 2014.

[6] C. Dicle, O. I. Camps, and M. Sznaier. "The way they move: Tracking multiple targets with similar appearance." In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2304-2311. 2013.
[7] H.B. Shitrit, J. Berclaz, F. Fleuret, and P. Fua. *Tracklet-based multi-commodity network flow for tracking multiple people*. No. EPFL-PATENT-186751. WO, 2013.
[8] B. R. Vatti, "A generic solution to polygon clipping." *Communications of the ACM* 35, no. 7 (1992): 56-63.
[9] A. Dehghan, S. M. Assari, and M. Shah. "Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4091-4099. 2015.
[10] M. Andriluka, S. Roth, and B. Schiele. "Monocular 3d pose estimation and tracking by detection." In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 623-630. IEEE, 2010.
[11] G. Shu, A. Dehghan, and M. Shah. "Improving an object detector and extracting regions using superpixels." In *Proceedings of the IEEE* G. R. Faulhaber, "Design of service systems with priority reservation," in *Conf. Rec. 1995 IEEE Int. Conf. Communications*, pp. 3–8. *Conference on Computer Vision and Pattern Recognition*, pp. 3721-3727. 2013.
[12] L. Wen, W. Li, J. Yan, Z. Lei, D. Yi, and S. Z. Li. "Multiple target tracking based on undirected hierarchical relation hypergraph." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1282-1289. 2014.
[13] S. Tang, B. Andres, M. Andriluka, and B. Schiele. "Subgraph decomposition for multi-target tracking." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5033-5041. 2015.
[14] A. Dehghan, Y. Tian, P. H.S Torr, and M. Shah. "Target identity-aware network flow for online multiple target tracking." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1146-1154. 2015.
[15] 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*, pp. 3682-3689. 2013.
[16] 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*, pp. 5537-5545. 2015.
[17] S. Wang and C. C. Fowlkes. "Learning Optimal Parameters for Multi-target Tracking." In *BMVC*, pp. 4-1. 2015.
[18] S. Wang and C. C. Fowlkes. "Learning Multi-target Tracking with Quadratic Object Interactions." *arXiv preprint arXiv: 1412.2066* (2014).
[19] K. Bernardin and Rainer Stiefelhagen. "Evaluating multiple object tracking performance: the CLEAR MOT metrics." *EURASIP Journal on Image and Video Processing* 2008, no. 1 (2008): 1-10..
[20] Wang, Bing, Li Wang, Bing Shuai, Zhen Zuo, Ting Liu, Kap Luk Chan, and Gang Wang. "Joint Learning of Convolutional Neural Networks and Temporally Constrained Metrics for Tracklet Association." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1-8. 2016.
[21] Y. Li, C. Huang, and R. Nevatia. "Learning to associate: Hybridboosted multi-target tracker for crowded scene." In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 2953-2960. IEEE, 2009.
[22] Y. Xiang, A. Alahi, and S. Savarese. "Learning to track: Online multi-object tracking by decision making." In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4705-4713. 2015
[23] C.-C. Chang and C.-J. Lin, LIBSVM: *a library for support vector machines. ACM Transactions on Intelligent Systems and Technology*, 2:27:1--27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

ISBN: 978-988-14047-4-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)