A Graph Transduction Game for Multi-target Tracking

Tewodros Mulugeta Dagnew*, Dalia Coppi†, Marcello Pelillo*, Rita Cucchiara†

*DAIS - Ca´ Foscari University
Venezia, Italy
Email: dagnew.tewodros@gmail.com, pelillo@dsi.unive.it

†DIEF - University of Modena and Reggio Emilia
Modena, Italy
Email: dalia.coppi@unimore.it, rita.cucchiara@unimore.it

arXiv:1806.07227v2 [cs.CV] 25 Jun 2018

Abstract—**Semi-supervised learning is a popular class of techniques to learn from labeled and unlabeled data. The paper proposes an application of a recently proposed approach of graph transduction that exploits game theoretic notions to the problem of multiple people tracking. Within the proposed framework, targets are considered as players of a multi-player non-cooperative game. The equilibria of the game is considered as a consistent labeling solution and thus an estimation of the target association in the sequence of frames. Patches of persons are extracted from the video frames using a HOG based detector and their similarity is modeled using distances among their covariance matrices. The solution we propose achieves satisfactory results on video surveillance datasets. The experiments show the robustness of the method even with a heavy unbalance between the number of labeled and unlabeled input patches.**

I. INTRODUCTION AND RELATED WORK

Semisupervised learning (SSL) has recently gained a considerable interest in the machine learning and computer vision communities. The reason of the diffusion of SSL method can be easily explained observing that in many real world applications unlabelled data are easy to find, while labelled data are expensive and difficult to retrieve. Typical examples can be found in digital forensics and video surveillance large amount of images and video are readily available, but, without effective automatic tools, corresponding annotation and analysis usually requires the human intervention. Data coming from surveillance cameras need to be examined in order to identify targets that may be captured by different cameras or in different time intervals, therefore in these contexts, many problems can be interpreted as SSL problems, where the initial labelled samples are available (*e.g* patches of the selected target, some initial frames of the video) in vast footage of unlabelled data.

A comprehensive survey on SSL can be found in [1] and, among the others, graph based algorithms have a relevant role. In such methods labeled and unlabelled input data are modeled as nodes of undirected graphs whose edges represent the similarity among data points. Labels for unlabeled instances are estimated by propagating the information available at labeled nodes to unlabeled nodes. The underline idea is that point close to each other (*i.e* point with high similarity) have the same label. Labels are usually estimated in regularization

frameworks where the aim is to minimize a function on the undirected graph [2], [3]. Graph based approaches have been effectively used to solve several problems in computer vision and pattern recognition [4], [5], [6], [7], [8], [9], [17] and also in medical imaging domain [10]. The main advantage of graph methods is to be non-parametric and transductive in nature (the solution is a cut of the graph, thus a label for unlabelled input point and not a general classification function as in the case of inductive methods [1]).

Referring to the video forensics and surveillance fields and considering the tracking problem in its common definition of "problem to estimate the location of target objects in a sequence of images starting from an initial detection", many different approaches have been proposed. In [11] a large experimental survey of various tracking approaches is presented evaluating the suitability of each approach in different situations and with different constraints (*e.g* assumptions on the background, on the motion model, on the occlusions, etc.). The survey also suggests how the problem of people tracking can be viewed as a data association problem, and in particular as a semi-supervised classification, where some examples of the people to track are initially detected.

The transductive learning paradigm has already been exploited to solve tracking and re-identification problems. The seminal work of [12] introduced the TL as a solution to the severe variation of the models in color tracking. They fitted the TL problem into an EM frameworks to estimate the pixel labels in hand and face color tracking. [13] proposed an on-line single target tracking using graph transduction applied to faces and cars. [14] proposed an on-line single target tracking and re-identification method based on a graph based formulation of the TL problem. They enforced the knowledge encoded in the labelled instances introducing an update strategy to avoid drift in the tracking and to allow re-identification in case of occlusions. In [15] , [16], [17], [18], [19] they consider all the pairwise relationships between detection responses in a temporal sliding window, which is used as an input to their optimization based on fully-connected edge-weighted graph.

In this paper we propose to exploit the graph transduction to track multiple people in videos. To our knowledge this is the first application of transductive learning to multiple target

tracking. Given few labelled patches of the people to follow we propose a formulation of the multitarget tracking as a graph transduction problem. We exploit the formulation of [20] based on a game theoretic framework and we prove its reliability in solving a real world problem. We propose an applications able to work off-line on pre-recorded video streams. Similarly to [14] we describe people patches with covariance matrices and we build the similarity graph using distance among covariance in Riemannian manifolds.

The rest of the paper is organized as follows: Section II summarizes the algorithmic framework proposed in [20] and subsequently Section III details how we propose to apply it in the context of multiple people tracking. In Section IV the performance of the application are evaluated on video surveillance datasets, finally Section V gives a glimpse on future works and possible extensions.

II. GRAPH TRANSDUCTION GAME

The theoretical formulation of the *Graph Transduction Game* (GTG) has been recently introduced in [20]. Starting from the basis of the transductive learning on undirected graph, they build a solution in which the label estimation is based on game-theoretic notions, in contrast to common solution based on the spectrum of the graph. Precisely the graph transduction is formulated as a non-cooperative multiplayer game and the labelling correspond to the Nash equilibria.

For notions about multi-player games we refer to [21], for completeness we only underline that the main idea is that a *game* is modelled as a strategic interaction among *players* where the goal of each player is to maximize its own *payoff* by choosing the best action (*pure strategy*) to play. The *graph transduction game* in [20] is formulated assuming that players $i \in \mathcal{I}$ participating in the game corresponds to a particular point in a data set $\mathcal{D} = \{d_1, \ldots, d_n\}$ and that the set of strategy among whom the players can choose is $S_i = \{1, \ldots, c\}$. Each strategy S_i expresses a certain hypothesis about its membership to a class and c is the total number of classes (*i.e* the the mixed strategy profile of each player $i \in \mathcal{I}$ lies in the c-dimensional simplex Δ_i).

Since the problem is a problem of SSL, the players belongs to two disjoint groups: those which already have knowledge of their membership, referred to as *labelled players* and denoted with the symbol $\mathcal{I}_\ell$, and those which do not have any idea about their membership at the beginning of the game, which are hence called *unlabelled players* and correspondingly denoted with $\mathcal{I}_u$. The labelled players $\mathcal{I}_\ell = \{\mathcal{I}_{\ell|1}, \ldots, \mathcal{I}_{\ell|c}\}$ do not need to maximise their payoff since they always play their already chosen k^{th} pure strategy where $k = 1, \ldots, c$. The transduction game can be easily reduced to a game with only unlabelled players $\mathcal{I}_u$ that need to find their mixed strategy $e_i^k \in \Delta_i$ and the fixed strategies of labelled players $\mathcal{I}_\ell$ act as bias over the choices of unlabelled players. For the Nash equilibrium theorem [22] the GTG always has equilibrium in mixed strategies that corresponds to a steady state where each player plays a strategy that could yield the highest payoff when the strategies of the remaining players are kept fixed, and it

provides us a globally consistent labeling of the data set. Once an equilibrium is reached, the label of a data point (player) i is simply given by the strategy with the highest probability in the equilibrium mixed strategy of player i as

$$\phi_i = \arg \max h = 1 \ldots c x_{ih}, \tag{1}$$

thereby yielding a crisp classification.

Similarly to other graph transduction methods the data are represented with an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ is the set of nodes representing both labelled and unlabelled points, and $\mathcal{E}$ are the edges weighted with an adjacency matrix $W = (w_{ij})$. Being the solution considered as the equilibrium in a non-cooperative game, the adjacency matrix W is used to compute the pay-off between players. And being the game considered as an instance belonging to the class of *polymatrix games* [23], [24] where players are nodes of a graph and every edge denote a two-player game between corresponding pair of players, the partial pay-off matrix between two players i and j is computed as $A_{ij} = w_{ij} \times I_c$ where I_c is the identity matrix of size c. The pay-offs are then computed as $u_i(e_i^h) = \sum_{j=1}^{n}(A_{ij}x_j)_h$ and $u_i(x) = \sum_{j=1}^{n} x_i^T A_{ij} x_j$.

The Nash equilibria, thus the labelling for unlabelled points, is computed using the *evolutionary approach* [25], [26]. The dynamic interpretation of Nash equilibria through the evolutionary approach imagines that the game is played repeatedly, generation after generation, during which a selection process acts on the multi-population of strategies, thereby resulting in the evolution of the fittest strategies. The particular class of dynamics used in this are the so called *imitation dynamics* given by

$$\dot{x}_{ih} = x_{ih} \left[\sum_{l \in S_i} x_{il} \Big(\phi_i \left[u_i \left(e_i^h - e_i^l, x_{-i} \right) \right] \right. \\ \left. - \phi_i \left[u_i \left(e_i^l - e_i^h, x_{-i} \right) \right] \Big) \right] \tag{2}$$

where the dot signifies derivative w.r.t. time and $\phi_i(u_i)$ is a strictly increasing function of u_i. The multi-population version of the replicator dynamics is obtained when ϕ_i is taken as the identity function, *i.e.* $\phi_i(u_i) = u_i$, as:

$$\dot{x}_{ih} = x_{ih} \left(u_i(e_i^h, x_{-i}) - u_i(x) \right) \tag{3}$$

[20] demonstrate how in both the discrete and continuous time version of the imitation dynamics the fixed points of Eq. 3 are Nash equilibria.

III. MULTITARGET TRACKING

This section describes the extension of the theoretical GTG framework presented in the previous section, to a solution for multiple people tracking and details the graph construction and the people extraction and description. An overview of the main steps of the proposal are depicted in Fig. 1.

We assume that the number of classes C is known and that all classes are present in the labelled data *i.e* we assume that the initial targets estimates are given. The labels y_i for the
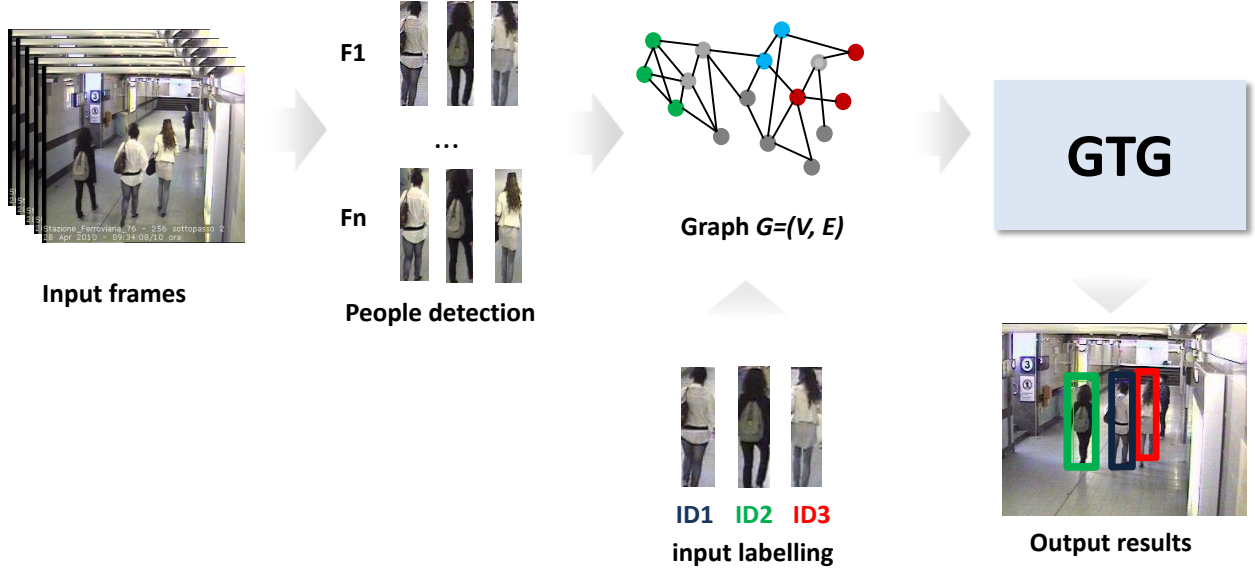
Fig. 1. Overview.

labelled points are in the set $y_i = 1, \dots, C$ and represent the *IDs* of the different targets.

Since graph-based learning methods enforce label smoothness over the graph, it is important that the graph is built exploiting a meaningful and robust similarity measure capable of capture people similitude in patches of the same target and differentiate between different targets. We decided to model people appearance using covariance matrix feature descriptors, [27]. This representation has been adopted in multiple approaches [14], [28], [29] because of its robustness in capturing shape, location and color information.

A. Covariance representation

Considering d different pixel features extracted from the image patches, the resulting covariance matrix C is a square symmetric matrix $d \times d$ where the diagonal entries represent the variance of each feature and the non-diagonal entries represent the correlations. We decided to model each pixel within a people patch with its position (x, y), its *HSV* intensity values and its derivatives information. The 9-dimensional resulting feature vector z_i is thus :

$$z_i = [\ x \ \ y \ \ H \ \ S \ \ V \ \ G_x \ \ G_y \ \ mag(x,y) \ \ o(x,y) \ ]^T \quad (4)$$

where G_x and G_y are the first order derivatives of the intensities calculated through Sobel operator w.r.t. x and y, $mag(x,y) = \sqrt{G_x^2 + G_y^2}$ and $o(x,y) = arctan\left(\frac{G_y}{G_x}\right)$ are respectively the magnitude and the angle of the first order derivatives. The covariance C of a patch of size $n = w \times h$ is then computed as $C = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \mu)(z_i - \mu)^T$ with μ the means of the vectors z_i. A schematic representation of the feature extraction is shown in Fig. 2. The choice of using
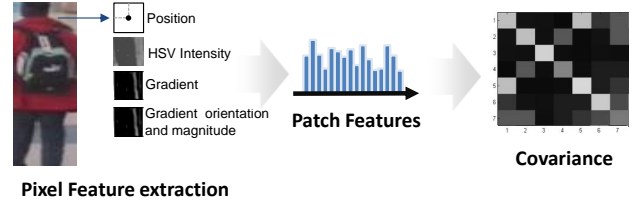

Fig. 2. Covariance matrix computation.

the HSV color space instead of the basic RGB is because of its invariance w.r.t. scale and shift variance of light intensity.

B. Graph construction

In order to build the payoff matrix A, the distances between covariance matrices are necessary, and, since they do not lie on the Euclidean space, we need to use an *ad-hoc* metric. Such distance metric between covariance matrices, proposed in [30] is equal to the sum of the squared logarithms of the generalized eigenvalues. Formally the distance between two matrices C_i and C_j is expressed as:

$$\rho(C_i, C_j) = \sqrt{\sum_{i=1}^{d} \ln^2 \lambda_k(C_i, C_j)} \quad (5)$$

where $\lambda_k(C_i, C_j)_{k=1\dots d}$ are the generalized eigenvalues. It is proven that ρ satisfies the metric axioms, positivity, symmetry, triangle inequality, for positive definite symmetric matrices.

Assuming that our system works off-line on pre-recorded videos, the adjacency matrix W is built over the complete set

Fig. 3. Examples of frames taken from THIS (left) and Caviar (right) datasets. Coloured bounding boxes show the obtained tracking results.
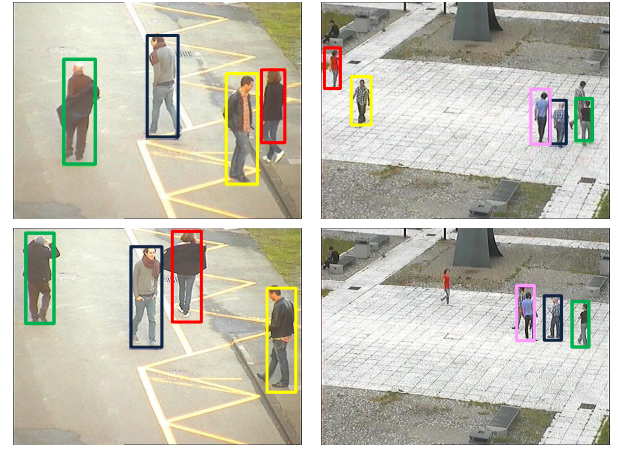

Fig. 4. Examples of frames taken from the 3dPes datasets. Coloured bounding boxes show the obtained tracking results.

of people patches extracted from all the images composing the video. The pay-offs are then specified in terms of the normalized similarity matrix $\widehat{W} = D^{-1/2} W D^{-1/2}$ with $D = (d_{ii})$ being the diagonal degree matrix of W whose elements are given by $d_{ii} = \sum_j w_{ij}$ since it has been demonstrated that the normalization yield to better performance than the original similarity.

In our particular configuration the players of the game are the different target to track and finding the Nash equilibria is equivalent to find the association of the targets in the set of images of the video stream. The

IV. EXPERIMENTS

In this section we report the results obtained with the two proposed configurations on-line and off-line for people tracking and association. We evaluate our proposal on a set of video sampled from three datasets, namely:

- **THIS** [1]: the videos we used in the evaluation are recorded along the platforms and underpasses of a train station and usually exhibit people walking in small groups or alone.
- **CAVIAR** [2]: clips of this dataset are collected in the hallway of a shopping center in Portugal.
- **3DPes** [3]: this dataset, [31], consists of videos collected at different times of the day in an university campus from different surveillance cameras. The main challenge of this dataset, originally proposed for re-identification tasks, is the large number of occlusions of the targets, we thus use it to evaluate the behaviour of our method in critical situation.

Fig. 3 and Fig. 4 show some frames taken respectively from **THIS** and **Caviar** and **3dPes**. As explained in Sec. III the GTG framework assumes that for each frame the patches containing the people on the scene are given, thus we initially

extracted those patches using a conventional people detector based on Histogram of Oriented Gradient. Working off-line on the people patches the problem can also be considered as one of multiple object data association, therefore we measured the performance in terms of mean object Precision and mean object Recall, where we considered a *True Positive* as a patch classified correctly with its label, a *False Positive* a misclassified patch (*e.g* the ith target label is assigned to a different person) and *False Negative* a missing estimation (*e.g* the ith target label is non assigned in the frame even if that target is present). In order to abstract the overall performance we also evaluate the F-measure.

Since the approach we proposed is off-line and relies on pre-recorded sequence, we evaluate the results varying the number of initially labelled frames. These frames were randomly chosen and the results have been averaged on 20 different executions. Table I reports the F-measure obtained using a initial labelling of five frames on the three datasets and summarizes the averaged length of the evaluated videos in number of frames and the number of targets to track for each dataset. Fig. 5(a) and 5(b) show the values of mean precision and recall on the three datasets.

Clearly the obtained results improve as we increase the labelled frames as illustrated in Fig. 5, but almost saturate to satisfactory values when the number of frames is fixed to 5 demonstrating good reliability without requiring a large number of labelled data. The results also show how increasing the number of labelled frames also increase the stability of the solution while labelling only one or three frames, despite the precision and recall reach adequate levels, the standard deviation is large. The best performing dataset is **THIS** reaching the 100 % of precision and accuracy, but results are good even on challenging dataset such as **3dPes**, where the number of target is higher and the targets are not always correctly identified by the detector due to occlusions, though on the **3dPes** dataset the standard deviation of the results is higher. We would like to highlight that the video length reported in Table I is an

---

[1] http://www.openvisor.org

[2] http://homepages.inf.ed.ac.uk/rbf/CAVIAR/caviar.htm

[3] http://www.openvisor.org/3dpes.asp

average on different sequences, but, especially with the **3dPes** dataset we stressed the algorithm increasing the number of frames from 200 to 430 and even with the highest number of frames the results were satisfactory with the F-measure of roughly the 90%. Few obtained tracking results are proposed in Fig. 3 and Fig. 4.

TABLE I
MULTITARGET TRACKING PERFORMANCES.

Dataset	# Frames	# Targets	F-measure
THIS	109	3	0.99
CAVIAR	140	4	0.96
3dPes	280	5	0.92

To further the evaluation we also compare our approach with the other tracking methods based on graph transduction. At this aim we recast our solution as one of single people tracking, labelling the input samples x_i with $y_i = +1$ when they correspond to the target and $y_i = -1$ otherwise. Results are reported in Tab. II, in particular we evaluate the GTG framework against the seminal Transductive Learning Tracker (TLT) of [13] and the work presented in [14], that proposed a single target Transduction Tracking (TT NoUpd) and also an improved version where the labelled data point are iteratively updated with an evolutionary spectral clustering (TT SpUpd). Even if we should specify that all the other methods work iteratively on-line on each frame of the video, our solution has been proven to be reliable even in a single target configuration and outperforms the other methods. The reason of the improvement over the state-of-the-art is justified both by the robustness of the GTG framework as already demonstrated in [20] for other classification tasks, and by the fact that working with graphs with more nodes increase the possible paths among nodes representing the target patches in different frames. In other words this overcomes the inevitable errors of the on-line methods when the people detection are imprecise in the evaluated frames.

TABLE II
COMPARISON WITH THE SINGLE TARGET TRACKING METHOD PROPOSED IN [14].

Dataset	Method	Precision	Recall	F-measure
THIS	*GTG*	1.00	0.85	0.92
	TLT	0.76	0.92	0.83
	TT NoUpd	0.80	0.93	0.86
	TT SpUpd	0.97	0.95	0.96
CAVIAR	*GTG*	0.90	0.92	0.91
	TT NoUpd	0.68	0.91	0.78
	TT SpUpd	0.78	0.90	0.84
	TLT	0.96	0.94	0.95
3dPes	*GTG*	0.87	0.99	0.93
	TT NoUpd	0.45	0.44	0.44
	TT SpUpd	0.63	0.66	0.64
	TLT	0.86	0.83	0.84

V. CONCLUSION

In this paper we proposed a method for multitarget people tracking based on a Graph Transduction Game framework. The GTG considers the different targets to track as players of a game, and estimates the missing labels by finding the Nash equilibria over the set of possible strategies, where the pay-offs of the different strategies are proportional to the similarity among people patches. We extracted people patches by using a HOG based detector and we represented their appearance with a covariance matrix.

The method has been evaluated over three different video surveillance datasets and obtained good results even with very few labelled initial frames. Our proposal also outperforms similar methods when considered in a single target configuration.

Given the promising results we obtained future work includes the extension to longer video sequence in order to test the robustness of the system also for long term tracking. We would like also to evaluate a on-line approach and insert mechanism to handle the insertion of new targets, or in other words to handle a variable number of classes C. Always in the framework of an on-line label estimation we would like to test the possibility of avoid the initial step of people detection, and instead, to use a sliding window approach to extract the patches, such as the method is independent from the performance of the detector.

REFERENCES

[1] X. Zhu, "Semi-supervised learning literature survey," 2006.
[2] T. Joachims, "Transductive learning via spectral graph partitioning," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2003, pp. 290–297.
[3] D. Zhou, J. Huang, and B. Schölkopf, "Learning from labeled and unlabeled data on a directed graph," in *Proceedings of the International Conference on Machine Learning (ICML)*, ser. ICML '05. New York, NY, USA: ACM, 2005, pp. 1036–1043.
[4] Y. T. Tesfaye, "Applications of a graph theoretic based clustering framework in computer vision and pattern recognition," 2018.
[5] E. Zemene, Y. T. Tesfaye, A. Prati, and M. Pelillo, "Simultaneous clustering and outlier detection using dominant sets," in *Pattern Recognition (ICPR), 2016 23rd International Conference on*. IEEE, 2016, pp. 2325–2330.
[6] E. Zemene, L. T. Alemu, and M. Pelillo, "Constrained dominant sets for retrieval," in *23rd International Conference on Pattern Recognition, ICPR 2016, Cancún, Mexico, December 4-8, 2016*, 2016, pp. 2568–2573.
[7] ——, "Dominant sets for "constrained" image segmentation," *CoRR*, vol. abs/1707.05309, 2017. [Online]. Available: http://arxiv.org/abs/1707.05309
[8] E. Zemene, Y. T. Tesfaye, H. Idrees, A. Prati, M. Pelillo, and M. Shah, "Large-scale image geo-localization using dominant sets," *arXiv preprint arXiv:1702.01238*, 2017.
[9] E. Zemene, S. R. Bulò, and M. Pelillo, "Dominant-set clustering using multiple affinity matrices," in *International Workshop on Similarity-Based Pattern Recognition*. Springer, 2015, pp. 186–198.
[10] T. M. Dagnew, L. Squarcina, M. W. Rivolta, P. Brambilla, and R. Sassi, "Learning from enhanced contextual similarity in brain imaging data for classification of schizophrenia," in *International Conference on Image Analysis and Processing*. Springer, 2017, pp. 265–275.
[11] A. W. M. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah, "Visual tracking: An experimental survey," *IEEE Trans. on Pattern Anal. Mach. Intell.*, vol. 99, 2013.
[12] Y. Wu and T. Huang, "Color tracking by transductive learning," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1. IEEE Computer Society, 2000, pp. 133–138.
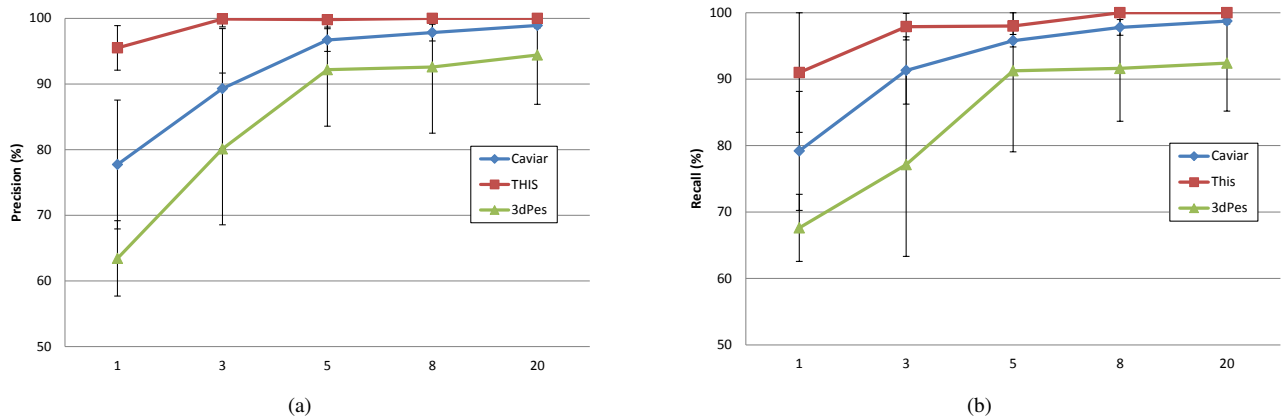
Fig. 5. Results reported in terms of Precision (a) and Recall (b) varying the labelled input frames. The number of labelled frames is reported on the horizontal axis.

[13] Y. Zha, Y. Yang, and D. Bi, "Graph-based transductive learning for robust visual tracking," *Pattern Recogn.*, vol. 43, no. 1, pp. 187 – 196, 2010.

[14] D. Coppi, S. Calderara, and R. Cucchiara, "People appearance tracing in video by spectral graph transduction," in *Proceedings of the IEEE International Conference on Computer Vision Workshops, (ICCV)*, Barcellona, Spain, Nov. 2011, pp. 920–927.

[15] Y. T. Tesfaye, E. Zemene, M. Pelillo, and A. Prati, "Multi-object tracking using dominant sets," *IET computer vision*, vol. 10, pp. 289–298, 2016.

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

[17] Y. T. Tesfaye, E. Zemene, A. Prati, M. Pelillo, and M. Shah, "Multi-target tracking in multiple non-overlapping cameras using constrained dominant sets," *arXiv preprint arXiv:1706.06196*, 2017.

[18] A. Dehghan, S. M. Assari, and M. Shah, "GMMCP tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4091–4099.

[19] Y. T. Tesfaye, "Multi-target tracking using dominant sets," M.Sc. thesis, Università Ca'Foscari Venezia, 2014.

[20] A. Erdem and M. Pelillo, "Graph transduction as a noncooperative game," *Neural Computation*, vol. 24, no. 3, pp. 700–723, 2012.

[21] J. W. Weibull, "Evolutionary game theory," 1995.

[22] J. Nash, "Non-Cooperative Games," *The Annals of Mathematics*, vol. 54, no. 2, pp. 286–295, Sep. 1951.

[23] L. Quintas, "A note on polymatrix games," *International Journal of Game Theory*, vol. 18, no. 3, pp. 261–272, 1989.

[24] J. T. Howson, "Equilibria of polymatrix games," *Management Science*, vol. 18, pp. 312–318, 1972.

[25] C. Daskalakis, P. W. Goldberg, and C. H. Papadimitriou, "The complexity of computing a nash equilibrium," in *Proceedings of the Thirty-eighth Annual ACM Symposium on Theory of Computing*, ser. STOC '06. ACM, 2006, pp. 71–78.

[26] C. Daskalakis, "On the complexity of approximating a nash equilibrium," 2011.

[27] F. Porikli, O. Tuzel, and P. Meer, "Covariance tracking using model update based on lie algebra," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, 2005, pp. 728–735.

[28] M. J. Metternich, M. Worring, and A. W. M. Smeulders, in *Transactions on data hiding and multimedia security V*, Y. Q. Shi, Ed. Springer-Verlag, 2010, ch. Color based tracing in real-life surveillance data, pp. 18–33.

[29] Y. Liu, G. Li, and Z. Shi, "Covariance tracking via geometric particle filtering," *EURASIP J. Adv. Signal Process*, pp. 1–22, 2010.

[30] W. Forstner, B. B. Moonen, and C. Gauss, "A metric for covariance matrices," 1999.

[31] D. Baltieri, R. Vezzani, and R. Cucchiara, "3dpes: 3d people dataset for surveillance and forensics," 2011.