Collaboration and spatialization for an efficient multi-person tracking via sparse representations

Loïc Fagot-Bouquet, Romaric Audigier, Yoann Dhome, Frédéric Lerasle

Collaboration and spatialization for an efficient multi-person tracking via sparse representations

Loïc Fagot-Bouquet[1], Romaric Audigier[1], Yoann Dhome[1], Frédéric Lerasle[2,3]
[1]CEA, LIST, Vision and Content Engineering Laboratory,
Point Courrier 173, F-91191 Gif-sur-Yvette, France
[2]CNRS, LAAS, 7, Avenue du Colonel Roche, F-31400 Toulouse, France
[3]Univ. de Toulouse, UPS, LAAS, F-31400 Toulouse, France
loic.fagot-bouquet@cea.fr

Abstract

Multi-person tracking is a very difficult problem in Computer Vision as a tracking algorithm is facing several issues, such as appearance changes, targets' occlusions and similar appearances between people. In an online tracking-by-detection algorithm, robust and discriminative specific appearance models help handling these difficulties. As done in single object tracking, we use sparse representations to extract local features of the targets and study how these representations can be specifically employed for multi-person tracking. Experiments on several datasets show that considering spatial information is crucial in order to improve the tracking performances with local descriptions compared to holistic features. Using large collaborative representations also improve the tracking results by naturally discarding irrelevant local patches.

1. Introduction

Multi-person tracking is a challenging topic, and particularly in video surveillance applications, affected by several issues that make this problem difficult. An efficient tracker has to deal with appearance variations, partial and severe targets' occlusions, and entrance-exit of people in the field of view of the camera.

This problem is either addressed offline [9, 13], using past and future frames, or online where only the past frames are considered to estimate the tracks of people [2, 10, 11, 16]. Recent online and offline trackers often rely on the tracking-by-detection scheme, using a classifier to detect pedestrian's locations in each frame.

Offline algorithms estimate the trajectories over a temporal window and can better handle the initialization and termination of the tracks compared to online methods. However, these methods can hardly include specific appearance and motion models of the targets, making difficult to differentiate people with similar appearances, and do not fulfill real-time applications.

In online approaches, detections are linked together frame to frame to reconstruct trajectories across time. Most of the time, specific appearance and motion models are learned online in order to find the best trajectories-detections assignment in a given frame. Therefore, employing robust and discriminative specific appearance models can significantly improve the tracking results of such methods [1, 11, 15].

Robust appearance models have been proposed for single object tracking in order to better cope with appearance changes and occlusions of the target. Sparse representations are largely employed in this field and used to describe the target with holistic or local features [5, 7, 8, 17]. A description based on local patches is argued to be more robust to occlusions, see Liu et al. in [7]. Jia et al. in [5] propose to include some spatial information within the sparse representations. This method appears to better localize the target location and naturally filter out irrelevant patches which can represent, for example, some occluded parts of the target.

We propose an online multi-person tracking-by-detection algorithm based on sparse representations of the targets described by local patches. We also use some spatial considerations, at the person or patch level, combined with collaborative representations among people or patches. Using collaborative representations with some spatial constraints, either at the person level or patch level, improves the tracking results and naturally determines relevant elements in the representations making them more discriminative. The best results on several public datasets are achieved when the most collaborative representations are employed with spatial considerations. In this setting, local descriptions yield better results than those obtained from holistic features of people.

This work is organized as follows. In Section 2, we present some related works and the tracking algorithm used in Section 3. We show the performances of the different choices of representations in Section 4 and the last section concludes this work.

2. Related work

While online multi-person approaches based on the tracking-by-detection paradigm were gaining popularity, some works began to consider more complex specific appearance models. Some employed local descriptions of the targets, like in [11] where part-based models inspired by DPM are used, or [15] where the motions of local patches on a regular grid are estimated in order to determine the most stable parts of the targets. In some methods, specific discriminative models are learned online to differentiate the related target from the surrounding ones [1, 11]. Existing approaches use these appearance models to estimate at each frame the affinity values for any track-detection pair. An Hungarian algorithm (or a greedy one) is then employed to find the best assignment between tracks and detections at a global level.

The description of the target appearance and the related model are a key element of single object tracking where a generic appearance of the target is unknown (as described in [12]). Sparse representations have been very popular in this field, and the related trackers employ a dictionary, composed by some views of the target, and assume that the target can be well approximated by a weighted linear combination of a few elements of this dictionary [8]. The dictionary can also include some elements from the background or the surroundings of the object in order to make the representations more discriminative [17]. Even though the majority of these approaches employ holistic descriptions of the target [6, 8, 17], some use a more local information by considering local patches [5, 7]. Particularly, the approach described in [5] uses local patches on a regular grid and includes some spatial considerations in order to penalize patches which are reconstructed with elements that do not share the same location on the grid. This naturally leads to select relevant patches that are reconstructed by elements from the same location and tends to discard patches from occluded parts.

Recently, a multi-person tracker was proposed in [10] using appearance models based on sparse representations. However, these appearance models are specific to each target and each of them is similar to a model used in a single object tracking approach. We consider that, in the context of a tracking-by-detection algorithm, appearance models should take in consideration the people jointly in order to be able to discriminate them efficiently. In fact, computing the affinity values between a given detection and all the possible tracks can be seen as a classification problem. Sparse representation can handle it, as it was proposed for face recognition, using a dictionary which combines some training examples of each class. The element to be classified is then represented in a collaborative way among all classes, and its class is estimated as the one that minimizes a residual reconstruction error [14]. In this paper, we propose to compute the affinity values for all the track-detection pairs using this kind of classifier with a local description of the targets, building on the work done in [4] which considers only holistic descriptions.

3. Proposed approach

We present in this section how sparse representations can be employed in a multi-person tracking system, using collaborative representations over local descriptions of the people and taking into account some spatial information between these descriptions.

3.1. System overview

Our system relies on the tracking-by-detection paradigm which means that a set of detections $\mathcal{D}$ is given at each frame by a pedestrian detector. These detections are associated with existing tracks, estimating an affinity value A_{ij} between the i^{th} track and the j^{th} detection in $\mathcal{D}$ with specific appearance and motion models. This association process consists in finding the best assignment between tracks and detections which is formulated as a maximum matching problem in a bipartite graph. This problem can be solved exactly with the Hungarian algorithm or with a greedy one for an approximated solution.

The trajectory handling is inspired from [16]. Tracks with a high association rate are declared confident, those with a low association rate are declared lost for a few frames before being definitively terminated.

The affinity values are estimated using sparse representations, but different choices are possible considering either the representations at the person level, as detailed in section 3.3, or at the patch level as detailed in section 3.4.

3.2. Dictionaries and sparse representations

The set of existing tracks at time t is denoted by $\mathcal{T} = \{T_1, ..., T_m\}$, and $\mathcal{D} = \{d_1, ..., d_l\}$ stands for the detections newly found. Each detection d_i is described by a set of p local features $\{y^1_{d_i}, ..., y^p_{d_i}\}$ related to patches selected on a regular grid or keypoints. For each track T_i, we combine the most recent features of the related target into a specific dictionary D_{T_i}. For any set of tracks $\mathcal{S} = \{T_{i_1}, .., T_{i_l}\}$, $D_{\mathcal{S}} = D_{T_{i_1}} \cup .. \cup D_{T_{i_l}}$ is called the joint dictionary of the tracks $T_{i_1}, ..., T_{i_l}$.

Given a feature vector y (related to some detection) and a dictionary D, a sparse code α^D_y can be defined by

$$\alpha^D_y = \underset{\alpha}{argmin}\ \frac{1}{2}||y - D\alpha||^2_2 + \lambda||\alpha||_1 \qquad (1)$$

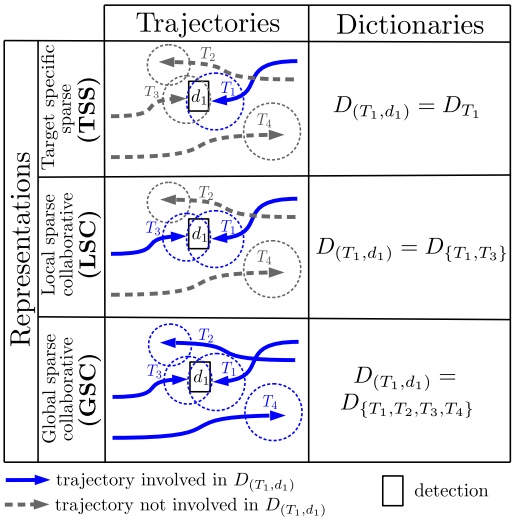

Figure 1. Sparse representations and related dictionaries.

where λ determines a trade-off between the reconstruction error $\frac{1}{2}||y - D\alpha||_2^2$ and the sparsity constraint $||\alpha||_1$.

$r(D, y, \alpha)$ stands for the reconstruction error of y with respect to the dictionary D and the code α, $r(D, y, \alpha) = \frac{1}{2}||y - D\alpha||_2^2$.

For any integer set I, the notation $\delta_I(\alpha)$ stands for the vector derived from α by setting to zero all the dimensions that are not in I. $r(D, y, \delta_I(\alpha))$ is called the residual reconstruction error for the set I.

3.3. Considerations at the person level

At each frame, we discard some associations between a track T_i and a detection d_j based on some spatial considerations. Specifically, we denote by $\mathcal{L}$ the set of all track-detection pairs (T_i, d_j) that can be linked together by considering two criteria, one based on the distance between d_j and the estimated location of T_i in the current frame, and the second one based on their shapes. $\mathcal{L}$ is defined as

$$\mathcal{L} = \{(T_i, d_j),\ \ dist_{T_i,d_j} < R_i \ \ and \ \ \frac{|h_i - h_j|}{h_i} < S_i\}$$

where $dist_{T_i,d_j}$ is the Euclidean distance between T_i and d_j, and h_j (resp. h_i) is the height related to d_j (resp. T_i). The values R_i and S_i are estimated for each track and are increasing when one is lost in order to allow a wider search area.

The affinity value A_{ij} between the track T_i and detection d_j is defined by

$$A_{ij} = -\sum_{k=1}^{p} r(D_{(T_i,d_j)}, y_{d_j}^k, \delta_{I_{T_i}}(\alpha_{y_{d_j}^k}^{D_{(T_i,d_j)}}))$$

if $(T_i, d_j) \in \mathcal{L}$, and $A_{ij} = -\infty$ otherwise. Each local feature $y_{d_j}^k$ from the detection d_j is represented with elements of a dictionary $D_{(T_i,d_j)}$ through its representation $\alpha_{y_{d_j}^k}^{D_{(T_i,d_j)}}$. $\delta_{I_{T_i}}$ sets to zero the coefficients in $\alpha_{y_{d_j}^k}^{D_{(T_i,d_j)}}$ not related to the track T_i and the affinity value A_{ij} is therefore derived from the residual errors of the local features $y_{d_j}^k$ for the track T_i.

We have several option for choosing the dictionary $D_{(T_i,d_j)}$, as proposed in [4], in order to take into account the spatial situation of the targets in the scene:

- A first possibility is to consider, for any track-detection pair (T_i, d_j),

$$D_{(T_i,d_j)} = D_{T_i}$$

This means that we consider the reconstruction error of each local feature of d_j over the specific dictionary related to the track T_i. This setting is called **TSS** (target specific sparse representations) as the sparse representation involved for estimating A_{ij} depends only on the specific dictionary D_{T_i} of T_i.

- The second possibility is to consider

$$D_{(T_i,d_j)} = D_{\{T_1,\ldots,T_m\}}$$

$D_{(T_i,d_j)}$ includes all the specific dictionaries, and the involved representations in this setting are collaborative among all the tracks. This setting is therefore called **GSC** (global sparse collaborative representation).

- The last option is

$$D_{(T_i,d_j)} = D_{\{T/(T,d_j)\in\mathcal{L}\}}$$

This time the representations involved are still collaborative but only among the tracks that are close to the considered detection. This setting is called **LSC** (local sparse collaborative representations).

These different possibilities are illustrated in Figure 1 where the trajectories involved in $D_{(T_i,d_j)}$ are specified.

3.4. Considerations at the patch level

In the previous section, the affinity values do not take into account any spatial information of the appearance of the target. By directly including all the local features into the dictionary we have lost all pertinent spatial information. However, we can use some spatial information by considering a dictionary $D_{(T_i,d_j)}^k$ and a function $\delta_{I_{T_i}}^k$ specific to each local feature $y_{d_j}^k$. The affinity value between any track T_i and detection d_j with $(T_i, d_j) \in \mathcal{L}$ is now defined by

$$A_{ij} = -\sum_{k=1}^{p} r(D_{(T_i,d_j)}^k, y_{d_j}^k, \delta_{I_{T_i}^k}(\alpha_{y_{d_j}^k}^{D_{(T_i,d_j)}^k}))$$

We have three alternatives for designing the dictionaries $D_{(T_i,d_j)}^k$ and functions $\delta_{I_{T_i}}^k$:

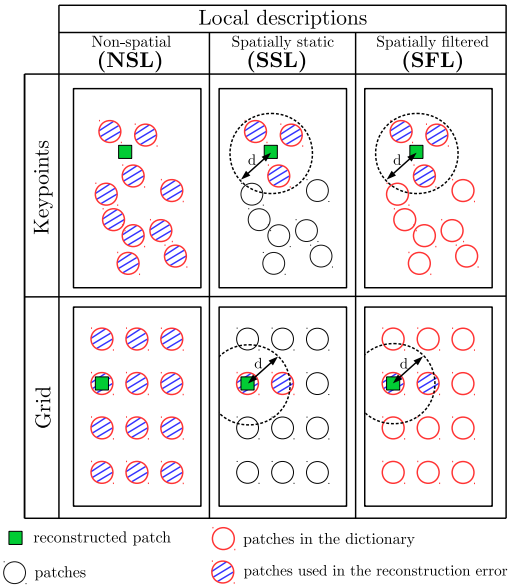

Figure 2. Local descriptions with some spatial information.

- The first one is to consider $D^k_{(T_i,d_j)} = D_{(T_i,d_j)}$ and $\delta^k_{I_{T_i}} = \delta_{I_{T_i}}$ as done in the last section. We will refer to this choice as **NSL** (non-spatial local description).

- A second possibility is directly inspired by [5]. In single object tracking, the method proposed in [5] shows that such an information can be easily used within the sparse representations and significantly improve the robustness of the tracking. In this paper, the target is described by local patches on a regular grid, and each patch is then reconstructed with a collaborative sparse representation among all the patches of the dictionary. However, the coefficients in the sparse code that are related to patches at a different location on the grid are set to zero. Therefore a patch has to be reconstructed mainly by patches from the same location in order to achieve a low reconstruction error. A patch which is occluded tends to be represented by patches at all possible locations and get a higher reconstruction error.

We adopt such strategy at the patch level, by choosing $D^k_{(T_i,d_j)} = D_{(T_i,d_j)}$ and considering for $\delta^k_{I_{T_i}}$ to set to zero all the coefficients that are related to local patches too distant from the location of $y^k_{d_j}$ or not related to the track T_i. In practice we consider a fixed threshold d and set to zero the coefficients associated to patches whose Euclidian distance to $y^k_{d_j}$ is above d. This strategy is called **SFL** (spatially filtered local description) in the experiments.

- A last possibility is to restrict the representations to use only patches that share the same location with the reconstructed one, using in $D^k_{(T_i,d_j)}$ only the elements of $D_{(T_i,d_j)}$ that are related to local patches near the location of $y^k_{d_j}$ and considering $\delta^k_{I_{T_i}} = \delta_{I_{T_i}}$. Contrary to the previous approach, the representations are no more collaborative between patches that do not share the same location. This strategy is called **SSL** (spatially static local description).

These possibilities are illustrated in Figure 2 where the patches used in $D^k_{(T_i,d_j)}$ and those considered in the reconstruction error (not set to zero by $\delta^k_{I_{T_i}}$) are specified.

4. Experimental evaluations

4.1. Implementation details

The proposed approach is implemented in C++ and is evaluated on single core at 2.7 GHz. As described in Figure 2, local patches are selected on a regular grid or interest points. We consider a grid of 3x4 patches or use an Harris corner detector to select 10 keypoints inside each detection bounding box (using non-maxima suppression), which means that $p = 10$ when using keypoints and $p = 12$ when using a regular grid. Each detection bounding box is resized to 64x128 pixels for finding local patches of size 32x32 pixels. We directly use the RGB intensity values of each local patch, resized to 16x16 pixels in order to reduce the dimensionality of the features.

We solve the assignment of the detections to tracks using a greedy algorithm, and use Kalman filters to predict the targets' locations in the next frame. For each variant we keep the same set of parameters for all datasets. Specific dictionaries are composed by the features from the 30 most recent views of the target, and the parameter λ in Eq. (1) is fixed to 0.1 (we got similar or slightly worse results with some other values).

Combining the different possibilities at the person level (**TSS**, **LSC** or **GSC**) with those at the patch level (**NSL**, **SSL** or **SFL**) and the two ways of selecting patches (using a grid or keypoints), 18 approaches are evaluated. We also compare our approaches using holistic features with the three settings described in section 3.3 and using the intensity values of the detections resized to 32x32 pixels.

4.2. Experimental setting

As not all state-of-the-art trackers provide results on the same datasets and because the related sets of detections are not always available, we use the following datasets to fairly compare our methods: PETS S2L1, PETS S2L2, TownCenter and ParkingLot. The same detections as [16] were used, which uses two sets of detections for the PETS and TownCenter datasets. The performances of other state-of-the-art online trackers [2, 10, 11, 16] are also indicated for these datasets and the same detections (when available, we use the trajectories from the authors' website or report them from the papers otherwise). These performances are evaluated with the CLEARMOT metrics [3] (composed of metrics like MOTA and MOTP) computed using the implementation from [16] with a standard overlap threshold of 0.5.

Description		TSS	LSC	GSC
Holistic		61.5	62.2	62.7
Keypoints	**NSL**	58.8	61.3	62.6
	SSL	60.0	61.8	63.1
	SFL	62.4	62.4	**63.4**
Grid	**NSL**	58.1	60.9	62.2
	SSL	59.7	61.6	62.9
	SFL	62.7	62.3	62.7

Table 1. Average MOTA values. Best value in bold and red, best values for each description underlined in blue.

Description		TSS	LSC	GSC
Holistic		129.0	127.5	125.7
Keypoints	**NSL**	157.5	118.1	107.2
	SSL	135.2	122.8	104.7
	SFL	116.4	107.4	**101.8**
Grid	**NSL**	172.7	137.1	114.8
	SSL	147.2	127.0	108.2
	SFL	119.8	110.8	107.5

Table 2. Average number of ID-Switches. Best value in bold and red, best values for each description underlined in blue.

4.3. Results analysis

The average MOTA scores for the different variants are shown in Table 1, and the average number of ID-Switches in Table 2. First of all, we can see that, despite their higher complexity, the methods with non-spatial local descriptions (**NSL**) do not produce better results than the holistic ones. However, imposing some spatial constraints (**SSL** and **SFL**) improves the tracking performances. Using local features with spatially filtered descriptions (**SFL**) improves the scores both in terms of MOTA and ID-Switches (still compared to holistic features). Using patches sampled on a regular grid produces lower results compared to patches around keypoints, and one can argue that using keypoints gives more stable and discriminant patches.

The most collaborative representations at the person level (**GSC**) yield better results (compared to **TSS** and **LSC**). At the patch level, the approach with spatial constraint and collaborative representations among patches (**SFL**) also yields superior results. The variant achieving the best results is the one which combines the most collaborative representations with spatial constraints (**GSC-SFL**).

As argued in [5], using collaborative representations among patches with spatial constraints naturally discards irrelevant patches since the related representations are more easily spread among all patches and not only among patches sharing the same location in the detection. When using collaborative representations among all tracks (**GSC**), we can assume that new people or false detections are associated to representations that are also more easily spread among all the tracks. Therefore the residual errors for the tracks which are close to irrelevant detections should be higher than those using local representations, avoiding some mismatches.

Data	Det.	Method	MOTA	IDS	MOTP	FP	MS
S2L1	**[9]**	**[16]**	69.9%	35	71.2%	805	557
		GSC-H	70.2%	25	65.6%	716	641
		GSC-K	**70.7%**	**19**	65.6%	732	606
	[16]	**[16]**	70.0%	21	71.7%	543	827
		GSC-H	71.1%	20	73.1%	461	857
		GSC-K	**72.5%**	**18**	73.1%	419	838
S2L2	**[9]**	**[16]**	43.1%	347	69.5%	1318	4189
		GSC-H	40.7%	230	66.0%	1553	4292
		GSC-K	**43.8%**	**177**	66.1%	1316	4263
	[16]	**[16]**	39.3%	287	69.0%	1416	4536
		GSC-H	**43.7%**	191	71.1%	1056	4526
		GSC-K	43.6%	**164**	71.2%	872	4743
Town Center	**[16]**	**[16]**	60.7%	212	71.2%	7295	20549
		GSC-H	**61.3%**	193	71.6%	3984	23472
		GSC-K	61.2%	**157**	71.6%	4053	23487
	[2]	**[16]**	63.4%	446	74.5%	9359	16302
		[2]	61.3%	318	80.5%	12309	14982
		GSC-H	66.0%	204	74.8%	6784	17286
		GSC-K	**66.6%**	**162**	74.8%	6636	17065
Parking Lot	**[11]**	**[10]***	84.5%	**4**	73.2%	-	-
		[11]*	79.3%	-	74.1%	-	-
		GSC-H	**85.6%**	17	71.3%	266	774
		GSC-K	85.4%	16	71.2%	287	771

Table 3. CLEARMOT metrics on various sequences (best values in bold and red for MOTA and ID-Switches (IDS), second best underlined in blue). The symbol * means the associated scores have been directly reported from the related papers. Det. specifies the set of detections used to feed the different trackers.

The CLEARMOT metrics for the most collaborative representations (**GSC** strategy), using either holistic features (**GSC-H**) or local patches with keypoints and spatially filtered descriptions (**GSC-K**) are shown in Table 3 and compared with other state-of-the-art online trackers. Our approaches yield better results in term of MOTA and ID-Switches on all the datasets except for ParkingLot where our methods still produce more ID-Switches. Using a local description (**GSC-K**) instead of a holistic one (**GSC-H**) gives better or very similar MOTA scores, and reduces significantly the number of ID-Switches. Some tracking results for the **GSC-K** approach are shown in Figure 3.

The main issue of the local descriptions is that the size of the involved dictionaries is larger (ten times larger with our parameters) than with holistic features. For this reason, the computational time becomes prohibitive. Optimizations based on active sets, as proposed in [4], allow the **GSC-H** approach to be real-time but are still insufficient in the case of the **GSC-K** approach.

Figure 3. Illustrations of some of our tracking results for the **GSC-K** approach.

5. Conclusion

In this work, we have investigated how sparse and collaborative representations of the targets could be used in a multi-person tracking application. We have shown that local descriptions improve the tracking results when some spatial constraint is combined with collaborative representations among local patches. An evaluation of our approach on several datasets shows the consistency of these observations and confirms that our approach is competitive when compared to other online tracking systems.

In future work, we could study how approximating sparse representations on large dictionaries in order to make the **GSC-K** approach real-time.

References

[1] S. Bae and K. Yoon. Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning. In *Conference on Computer Vision and Pattern Recognition*, pages 1218–1225. IEEE, 2014.

[2] B. Benfold and I. Reid. Stable multi-target tracking in real-time surveillance video. In *Conference on Computer Vision and Pattern Recognition*, pages 3457–3464. IEEE, 2011.

[3] K. Bernardin and R. Stiefelhagen. Evaluating multiple object tracking performance: The clear mot metrics. In *EURASIP J. Image and Video Processing*, volume 2008, 2008.

[4] L. Fagot-Bouquet, Y. Dhome, R. Audigier, and F. Lerasle. Online multi-person tracking based on global sparse collaborative representations. In *International Conference on Image Processing*. IEEE, 2015.

[5] X. Jia, H. Lu, and M. Yang. Visual tracking via adaptive structural local sparse appearance model. In *Conference on Computer Vision and Pattern Recognition*. IEEE, 2012.

[6] H. Li, C. Shen, and Q. Shi. Real-time visual tracking using compressive sensing. In *Conference on Computer Vision and Pattern Recognition*. IEEE, 2011.

[7] B. Liu, J. Huang, L. Yang, and C. Kulikowsk. Robust tracking using local sparse appearance model and k-selection. In *Conference on Computer Vision and Pattern Recognition*, pages 1313–1320. IEEE, 2011.

[8] X. Mei and H. Ling. Robust visual tracking and vehicle classification via sparse representation. In *Transactions on Pattern Analysis and Machine Intelligence*, volume 33, pages 2259–2272. IEEE, 2011.

[9] A. Milan, S. Roth, and K. Schindler. Continuous energy minimization for multitarget tracking. In *Transactions on Pattern Analysis and Machine Intelligence*, volume 36, pages 58–72. IEEE, 2014.

[10] M. A. Naiel, M. O. Ahmad, M. Swamy, Y. Wu, and M. Yang. Online multi-person tracking via robust collaborative model. In *International Conference on Image Processing*. IEEE, 2014.

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

[12] A. W. M. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah. Visual tracking: An experimental survey. In *Transactions on Pattern Analysis and Machine Intelligence*, volume 36, pages 1442–1468. IEEE, 2014.

[13] X. Wang, E. Turetken, F. Fleuret, and P. Fua. Tracking interacting objects optimally using integer programming. In *European Conference on Computer Vision (ECCV)*, pages 17–32. IEEE, 2014.

[14] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. In *Transactions on Pattern Analysis and Machine Intelligence*, volume 31, pages 210–227. IEEE, 2009.

[15] Z. Wu, J. Zhang, and M. Betke. Online motion agreement tracking. In *British Machine Vision Conference*. BMVA Press, 2013.

[16] J. Zhang, L. L. Presti, and S. Sclaroff. Online multi-person tracking by tracker hierarchy. In *Advanced Video and Signal-Based Surveillance*, pages 379–385. IEEE, 2012.

[17] W. Zhong, H. Lu, and M. Yang. Robust object tracking via sparsity-based collaborative model. In *Conference on Computer Vision and Pattern Recognition*. IEEE, 2012.