MULTI-OBJECT TRACKING USING SPARSE REPRESENTATION

Weizhi Lu, Cong Bai, Kidiyo Kpalma and Joseph Ronsin

Université Européenne de Bretagne, France
INSA de Rennes, IETR, UMR 6164, F-35708, RENNES

ABSTRACT

Recently sparse representation has been successfully applied to single object tracking by observing the reconstruction error of candidate object with sparse representation. In practice, sparse representation also shows competitive performance on multi-class classification, and thus is potential for multi-object tracking. In this paper we explore this technique for on-line multi-object tracking through a simple tracking-by-detection scheme, with background subtraction for object detection and sparse representation for object recognition. Final experiments demonstrate that the proposed approach only combining color histogram and 2-dimensional coordinates as features, achieves favorable performance over state-of-the-art work in persistent identity tracking.

Index Terms— multi-object, tracking, sparse representation

1. INTRODUCTION

Multi-object tracking is a technique that locates and recognizes a number of objects in some sequential video frames. Compared to single object tracking, it presents more challenges on objects discrimination. As a novel technique for classification, sparse representation has shown promising performance on face recognition [1, 2]. In this paper, we are motivated to further explore its potential of classification for multi-object tracking.

Recently sparse representation has been successfully applied to single object tracking [3–11] by locating the candidate object with minimal reconstruction error through the sparse representation of templates. Unfortunately, these works only use sparse representation for object representation, while ignoring its potential on multi-class classification [1, 2]. Compared to traditional classifiers, like SVM [12], sparse representation-based classification (SRC) is also competitive in computation. For instance, the recognition of object can be implemented by solving one l_1-regularization problem for SRC. Conversely, to obtain better performance, SVM usually has to divide multi-class classification problem into multiple binary classification problems. Especially, the number of binary classifiers increases exponentially with the number of objects. When new samples appear, each binary classifier is required to train subspaces again by solving a l_1 or l_2 regularization problem.

Here we take the popular tracking-by-detection scheme [13–16] for experiments: objects are first detected by background subtraction and then discriminated by SRC. For static scenes, background subtraction [17–19] is still a good option since it can obtain more reliable results compared to popular body detection methods preferred by dynamic scenes [20, 21]. Since objects can be located previously by background subtraction, for simplicity, we are allowed to not use special motion estimation algorithms at some cost of tracking fluency. In this paper, to better discriminate objects sharing similar appearances, object is represented with a vector combining color histogram as well as the 2-dimensional coordinates of object center.

The rest of this paper is organized as follows. In section 2, SRC is studied in terms of two types of solution algorithms. In section 3, multi-object tracking approach with SRC is detailed. In section 4, experimental results are described. Finally, a conclusion closes this paper.

2. OBJECT RECOGNITION USING SPARSE REPRESENTATION

2.1. Sparse representation-based classification

SRC in tracking can be described as the following problem. Let vector $\mathbf{y} \in \mathrm{R}^{m\times 1}$ denote one test object detected from current frame, and matrix $\mathrm{X} \in \mathrm{R}^{m\times n}$ be the database consisting of n labeled objects collected from former frames. Then test object can be approximated by the linear combination of known objects as

$$\mathbf{y} = \mathbf{X}\beta + \epsilon \tag{1}$$

where coefficient β is required to be a k sparse vector, namely β only has $k \ll n$ nonzero entries; and ϵ is a tolerated error. For a clearer expression, $\mathrm{X} = [X_{G_1}, X_{G_2}, \ldots, X_{G_N}]$ is further segmented into N sub-matrices respectively corresponding to N labeled classes of objects. And each class includes n_i objects, $X_{G_i} = [x_{i_1}, x_{i_2}, \ldots, x_{i_{n_i}}]$, where $1 \leqslant i \leqslant N$ and $\sum_{i=1}^{N} n_i = n$. Finally, test object is classified into the class

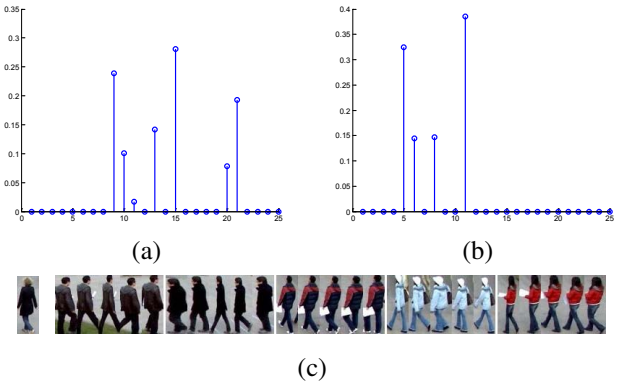


Fig. 1: One novel test object (left) in (c) is linearly approximated by 25 labeled training samples in (c), which sequentially correspond to 5 classes of objects in database. Sparse coefficients by LARS [22] in (a) and by group-OMP [23] in (b) both scatter into 3 classes.

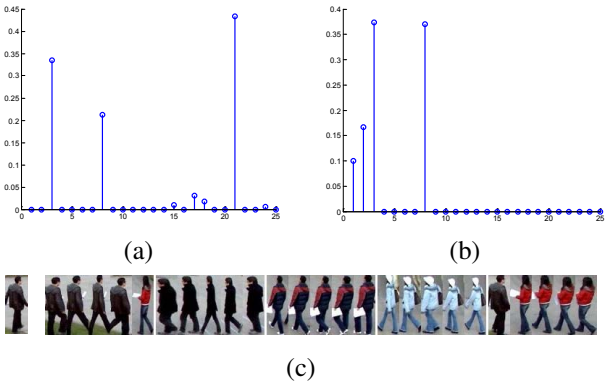


Fig. 2: One test object (left) in (c) is linearly approximated by 25 labeled training samples in (c), which sequentially correspond to 5 classes of objects in database. *class* 1 and *class* 5 both have a false training sample. According to sparse coefficients, the test sample belonging to *class* 1 is recognized successfully by group-OMP [23] in (b), and incorrectly determined as *class* 5 by LARS [22] in (a).

X_{G_i} satisfying

$$\min_{X_{G_i}} ||\mathbf{y} - X_{G_i}\delta_i(\beta)||_2, \quad 1 \leq i \leq N \tag{2}$$

where $\delta_i(\beta)$ is a function that sets all elements of vector β to zero except those corresponding to submatrix X_{G_i} . It is easy to understand that when columns of $\mathbf{X}$ are normalized, maximum entry in β usually corresponds to the most similar object. So for simplicity, $\text{argmax}_i\{\beta_i\}$ or $\text{argmax}_i\{||\delta_i(\beta)||_1\}$ is usually used to define the most similar class. In addition, with the goal of discrimination instead of a precise representation, we do not need exploring additional trivial template, like Gaussian matrix or identity matrix, to approximate noise or occlusion.

It is worth mentioning a special case when test object is novel and out of database, the nonzero entries of β tend to scatter among classes rather than focus on some single class as Figure 1 shows. So the novel object can be defined, if the following condition is verified

$$\max\{\beta_i\} < \gamma \sum_{j=1}^{n} \beta_j$$

where $0 < \gamma < 1$ is a constant.

2.2. Solution algorithms

Clearly, the kernel of SRC is to derive sparse solution β from formula (1). This solution is usually formulated as a linear regression problem with l_1 penalty

$$\hat{\beta} = \text{argmin}\{\|y - \mathrm{X}\beta\|_2 + \lambda\|\beta\|_1\} \tag{3}$$

where λ is a penalty parameter. As a convex problem, it has been widely studied for variable selection and model estimation in statistics, and a number of algorithms like interior point, OMP [24] and LARS [22], are also proposed in the past decade. These algorithms all perform the greedy pursuit process, in which columns of $\mathbf{X}$ are selected one by one with respect to minimize the residual of y as well as the number of nonzero entries of β. In this case SRC is similar to the nearest neighbor classifier (NNC) which tries to search the nearest training sample for test sample. Recently, the group-based greedy pursuit algorithms like Group-LARS [25] and Group-OMP [23] are sequentially proposed with the form

$$\hat{\beta} = \text{argmin}\{\|y - \sum_{i=1}^{N} X_{G_i}\beta_{G_i}\|_2 + \lambda \sum_{i=1}^{N} \|\beta_{G_i}\|_2\} \tag{4}$$

in which columns of $\mathbf{X}$ are operated as group in each greedy pursuit step. In this case SRC performs like a nearest class classifier (NCC). Generally, group-based algorithms seem more robust for database with burst error as the example in Figure 2. However, our experiments show that the algorithms without group constraint are more suitable for tracking system since object is usually similar between two adjacent frames while suffering from great variation across some frames, and thus the similarity measurement between test object and a group of training samples is unreliable.

3. PROPOSED TRACKING SCHEME

3.1. Object detection and representation

For static camera without dense scenes, background subtraction is efficient for body area detection. Here in terms of computation cost, we still use traditional method [17] based on statistically modeling and pixel-wise subtraction instead of these complex methods with tiny performance gain [18,

(a) (b)

Fig. 3: Tracking results of SRC only with color feature (a) and SRC with feature integrating color and 2-dimensional coordinates (b). In (a), *object* 6 switches into *object* 2 after 13 frames due to similar appearance.

19]. The body area is customarily represented by two cascaded RGB histograms corresponding to upper body and lower body. Furthermore, to discriminate objects sharing similar appearance, the location information, normalized 2-dimensional coordinates of object center, is applied to represent object by concatenating it with normalized RGB vector. Its advantage is simply exemplified in Figure 3. And the weight ratio between location and color feature is tuned empirically.

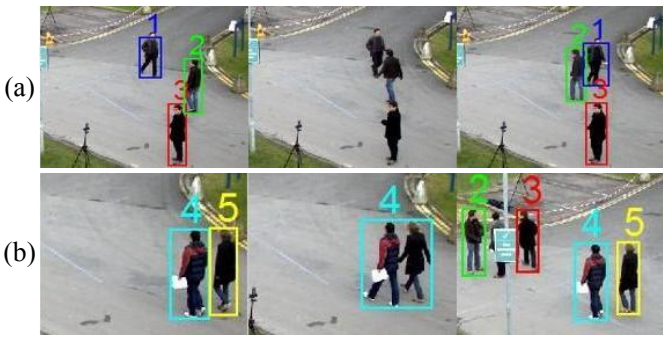
3.2. Overlapping

With background subtraction, the overlapping objects are subject to be detected as one object. And objects with tiny overlap usually can be detected by obvious size variation as shown in Figure 4(a). To avoid false samples updating or novel object definition in database, detected overlap is not processed and nor labeled in proposed scheme. As for the undetected overlap as Figure 4 (b) shows, it will be recognized as a unique object. Theoretically, the overlap tends to be linearly approximated by the objects it includes during sparse representation. Thus, the overlap is likely to be defined as the larger object it includes in our experiments. In this sense, SRC naturally avoids unwelcome novel object definition caused by overlapping.

3.3. Online database updating

Online database updating attempts to store and train recently recognized samples which are expected to be most similar to incoming test objects. This is critical for object recognition in tracking, since objects usually suffer from serious variation over time. The matrix structure of SRC is suitable for online database updating by renewing the columns of $\mathbf{X}$ frame by frame. Furthermore, as a multi-class classifier, there is no additional decision threshold training like SVM when new training sample is added. To enhance recognition rates and avoid false samples accumulation in database, we further impose some constraints on the scheme:

- to avoid identity switch, the initial detected object sample is always stored in database.
- at the beginning of experiments, the database is expanded by perturbing the initial sample with small Gaussian noise .



(a)

(b)

Fig. 4: Obvious overlap detected in (a) is not processed. And undetected overlap in (b) is usually recognized as the object at the forefront.

- the training samples of the same class share the same location information from the most recently updated sample.
- to avoid false recognition or novel object initialization, detected overlap is not updated.

Note that the popular database learning algorithms for reducing recovery error [7, 26] are not employed here since our system aims to exploit and evaluate the discriminability of SRC instead of recovery error.

4. EXPERIMENTS

There are few benchmark videos for multi-object tracking, especially the videos with few overlapped scenes that background subtraction can deal with. For comparison, we start experiment with a classic video from PETS'09, which has been evaluated by two state-of-the-art works, ETHZ [14] and EPFL [16]. ETHZ implements a robust tracking-by-detection scheme by combing body detection and particle filter. EPFL attempts to explore object appearance from a global view with multiple calibrated cameras. To further verify our performance, we also evaluate proposed approach on some sequences from PETS'06. (For video results, please refer to `http://youtu.be/SLyABs_nJeg`.)

Multi-object tracking usually faces three challenges: object switch during overlapping, new object initialization and re-recognition of re-entering objects. In the following part, we will briefly introduce two videos and then discuss the results in terms of aforementioned challenges.

4.1. Database PETS'09.

This video with 795 frames is recorded in a campus at 7 fps from a high view point. Ten persons walk in and out of scene, and some of them are similar in color. So it is a challenge for recognition by appearance. In the following comparison, they are recalled by the sequence number corresponding to their entries into scene. In our result, ten persons are labeled

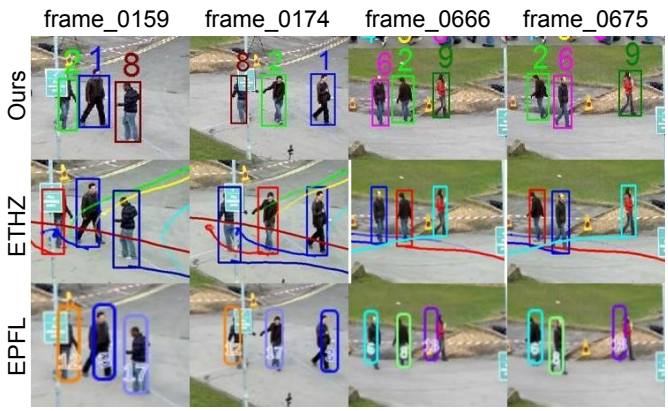


Fig. 5: Two examples on identity switch caused by overlapping. EPFL switches the identities of *objects* 12 and 17 in the first two frames, and exchange *objects* 6 and 8 in the last two frames. Conversely, the proposed approach and ETHZ work well.

with a number, and their initial samples are displaced on the top of each frame, as Figure 6 shows. In ETHZ and in EPFL, they are discriminated separately with color-box and number.

Results. Figure 5 illustrates two examples about identity switch between objects of similar appearances. In fact, the proposed approach shows better performance for discriminating objects on the whole video. This mainly benefits from the features involving object center coordinates. In proposed approach, one special case is that objects 4 and 5 are labeled together as object 4 for a long time due to overlap as shown in Figure 4(b). However, object 5 can be successfully recovered when they separate. In Figure 6, we give one example about object re-recognition and novel object initialization, in which proposed approach works well while other two methods fail. In fact, proposed approach shows best performance for persistent identity tracking in the whole video, as confirmed in Table 1. Otherwise, it should be recalled that proposed tracking process is not very fluent due to object detection failure as well as the lack of motion estimation.

4.2. Database PETS'06

We select a relatively crowded scene from S7.T6.B4 (frame _01685 to frame_01985), in which 11 objects suffer from serious size variation and illumination variation, and some of them also share similar color. For example, *objects* 4, 5 and 6 are hard to be distinguished by naked eyes when they walk away.

Results. As Figure 7 shows, the proposed approach successfully detects and initializes these 11 objects. Furthermore, there is no identity switch caused by occlusion or incorrect object initialization. This result further proves the robustness of SRC for multi-object classification.

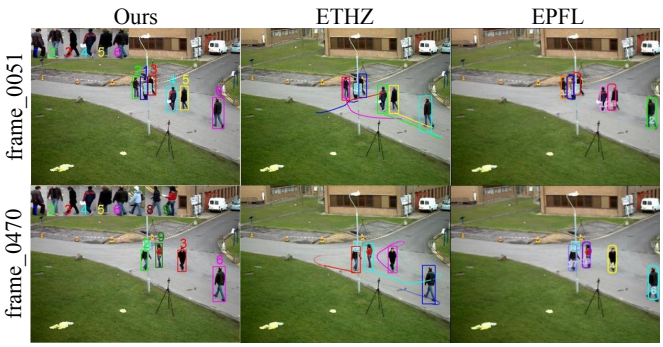


Fig. 6: Examples on object initialization and re-recognition. In the frame_0470, *object* 6 re-enters and *object* 9 first enters (referring our labels). *Object* 6 is not recovered in ETHZ and EPFL. And *object* 9 is incorrectly initialized to *object* 6 in ETHZ. In contrast, the proposed approach performs well on above two cases.

Table 1: Correct occurrences for objects entering or re-entering in scene (PETS'09). Value 0 indicates false object initialization.

Objects	1	2	3	4	5	6	7	8	9	10
Num. of entries	4	2	2	2	2	2	1	1	2	1
Ours	**4**	**2**	**2**	1	1	**2**	**1**	**1**	**2**	**1**
ETHZ	4	2	2	1	1	1	1	1	0	0
EPFL	2	1	1	1	1	1	1	1	2	1

Fig. 7: Tracking results for PETS'06. The bottom displays 11 objects that we initialize and track successfully. Our results have no identity switch or incorrect *novel* object initialization.

5. CONCLUSION

This paper has explored the potential of SRC for multi-object tracking through a simple yet effective tracking-by-detection scheme. By simply combining background subtraction for object detection and SRC for object recognition, the performance better than state-of-the-art is obtained in persistent identity tracking. As a multi-class classifier, SRC also shows advantage on complexity since it does not need training special classifier for each tracker during online database updating. In future, SRC will be developed for more advanced tracking schemes.

6. REFERENCES

[1] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Tran. PAMI*, vol. 31, pp. 210–227, 2009.

[2] M. Yang and L. Zhang, "Gabor feature based sparse representation for face recognition with gabor occlusion dictionary," *in ECCV*, 2010.

[3] X. Mei and H. Ling, "Robust visual tracking using l_1 minimization," *in ICCV*, 2009.

[4] H. Li, C. Shen, and Q. Shi, "Real-time visual tracking using compressive sensing," *in CVPR*, 2011.

[5] Q. Wang, F. Chen, W. Xu, and M.H. Yang, "Online discriminative object tracking with local sparse representation," *in WACV*, 2012.

[6] B. Liu, L. Yang, J. Huang, P. Meer, L. Gong, and C. Kulikowski, "Robust and fast collaborative tracking with two stage sparse optimization," *in CVPR*, 2010.

[7] B. Liu, J. Huang, L. Yang, and C. Kulikowsk, "Robust tracking using local sparse appearance model and *K*-selection," *in CVPR*, 2011.

[8] X. Mei, H. Ling, Y. Wu, E. Blasch, and L. Bai, "Minimum error bounded efficient l_1 tracker with occlusion detection," *in CVPR*, 2011.

[9] X. Jia, H. Lu, and M.H. Yang, "Visual tracking via adaptive structural local sparse appearance model," *in CVPR*, 2012.

[10] W. Zhong, H. Lu, and M.H. Yang, "Robust object tracking via sparsity-based collaborative model," *in CVPR*, 2012.

[11] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja, "Robust visual tracking via multi-task sparse learning," *in CVPR*, 2012.

[12] D.A. Ross, J. Lim, R.S. Lin, and M.H. Yang, "Incremental learning for robust visual tracking," *IJCV*, vol. 77, pp. 125–141, 2008.

[13] M. Andriluka, S. Roth, and B. Schiele, "People-tracking-by-detection and people-detection-by-tracking," *in CVPR*, 2008.

[14] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. V. Gool, "Online multi-person tracking-by-detection from a sing and uncalibrated camera," *IEEE Trans. PAMI*, vol. 33, no. 9, pp. 1820–1833, 2011.

[15] Z. Kalal, K. Mikolajczyk, and J. Matas, "Face-tld: Tracking-learning-detection applied to faces," *in ICIP*, 2010.

[16] H. B. Shitrit, J. Berclaz, F. Fleuret, and P. Fua, "Tracking multiple people under global appearance constraints," *in ICCV*, 2011.

[17] T. Horprasert, D.Harwood, , and L. S. Davis, "A statistical approach for real-time robust background subtraction and shadow detection," *in ICCV'99 Frame Rate Workshop*, 1999.

[18] H. Jiang, S. Fels, , and J. Little, "Adaptive background mixture models for real-time tracking," *In CVPR*, 1999.

[19] A. Amato, M. Mozerov, F. X. Roca, and J. Gonzlez, "Robust real-time background subtraction based on local neighborhood patterns," *EURASIP J. Adv. Sig. Proc.*, 2010.

[20] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *in CVPR*, 2005.

[21] B. Leibe, A. Leonardis, and B. Schiele, "Robust object detection with interleaved categorization and segmentation," *IJCV*, vol. 77, no. 1-3, pp. 259–289, 2008.

[22] B. Efron, T. Hastie, and R. Tibshirani, "Least angle regression," *Annals of Statistics*, vol. 32, pp. 407–499, 2004.

[23] A. C. Lozano, G. Swirszcz, and N. Abe, "Group orthogonal matching pursuit for variable selection and prediction," *in NIPS*, 2009.

[24] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," *in ACSSC*, 1993.

[25] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society and Series B and Methodological*, vol. 68, pp. 49–67, 2006.

[26] M. Aharon, M. Elad, and A. Bruckestein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Sign. Proc.*, vol. 50, no. 11, pp. 4311–4322, 2006.