Evaluation of Visual Tracking in Extremely Low Frame Rate Wide Area Motion Imagery

Haibin Ling[1], Yi Wu[1], Erik Blasch[2], Genshe Chen, Haitao Lang[1], and Li Bai[3]
[1]Computer and Information Science Department, Temple University, Philadelphia, PA U.S.A.
[2]Defense R&D Canada, QC, Canada
[3]Electrical and Computer Engineering Department, Temple University, Philadelphia, PA U.S.A.
hbling@temple.edu, wuyi@temple.edu, erik.blasch@gmail.com, genshe.chen@ieee.org,
haitaolang@hotmail.com, lbai@temple.edu

Abstract* - *Visual tracking in wide area motion imagery (WAMI) is an important problem in security related tasks. The extremely low frame rate and the large camera motion in such videos, however, introduce challenging constraints that distinguish the task from traditional image tracking. In this study we evaluate the performance of several state-of-the-art visual trackers on the wide area surveillance videos. Specifically, we compare five visual trackers on sequences selected from the Columbus Large Image Format (CLIF) dataset for the vehicle tracking task. The experiments are conducted in two configurations: one with background registration to compensate for camera motion, and the other without. We evaluate the tracking performances both qualitatively and quantitatively. The experimental results show that (1) traditional visual trackers meet problems in the wide area surveillance videos and (2) background registration helps enhance the tracking performances, although there exists needed improvement for operational robust performance. We expect the visual tracking evaluation on low-frame rate wide area surveillance videos to motivate future research to addresses the related challenges and provide annotated images for benchmark purposes.

Keywords: Visual tracking, registration, low-frame-rate, wide area surveillance.

1 Introduction

Visual tracking is a critical task in many security related applications such as surveillance, robotics, and human computer interaction. There are many factors that make visual tracking a challenging task, including occlusions, presence of noise, varying viewpoints, background clutter, pose change [26][16], and illumination changes [28]. Over decades, many algorithms have been proposed for visual tracking to addresses these challenges including: target representation [2][7][19], tracking inference [2], state propagation [21][22][9][13][27], simultaneous tracking and recognition, classification, and identification [5][6][29]. A thorough survey for general visual tracking can be found in [28].

A fundamental assumption in most previous studies is small and continuous variation in target movement and sensor motion. The small movement variation assumption naturally reflects how real time video sequences approximate human perceptions of real world motions. The assumption facilitates the modeling of target trajectories and thereafter guides the predication and propagation of tracking states. Such an assumption is, unfortunately, often violated in wide area surveillance videos which have been widely used to provide global security information

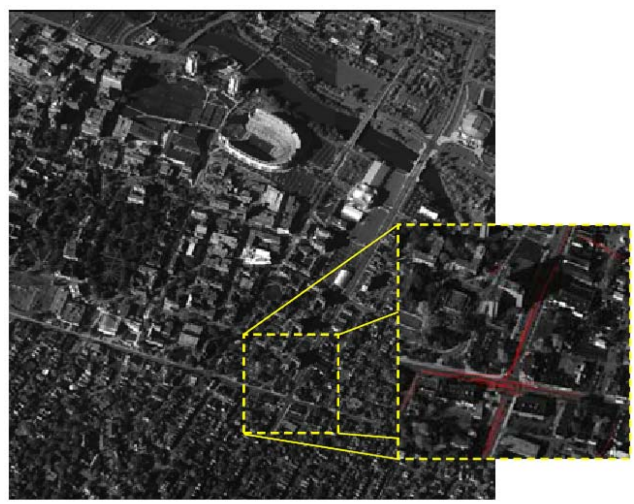


Figure 1. The AFRL CLIF dataset [1]. The illustration shows example system tracks in the zoom in window.

In this paper we investigate the performance of state-of-the-art visual trackers on wide area surveillance sequences, for the task of vehicle tracking in urban environment. Using the tracking outputs, we desire to provide moving intelligence (MOVINT) for security operators. For this purpose, we use the AFRL CLIF dataset [1], which contains wide area surveillance videos collected using six sensors (Figure 1). The collection included building mounted electro-optical cameras, millimeter wave infrared (MMWIR), Large Area Image Recorder (LAIR), and an electro-optical camera on an unmanned air vehicle (UAV). There are two main properties that make such dataset much more challenging than those used in traditional visual tracking tasks. First, the extremely low frame rate, two frames per second for the CLIF dataset, makes the trajectory prediction very difficult. Second, the fast camera motion, which arises from the aircraft flying, complicates background modeling. In addition, the tracking task is often made even more challenging due to other reasons including severe occlusions, similarity in background clutter, shadow, and other environmental effects. Detailed descriptions and illustrations of these challenges are discussed in Section 2. It is worth noting that the CLIF dataset has been previously used for evaluating registration methods in [18], and for designing criterion for multiple-target tracking in [15].

Many state-of-the-art visual trackers perform accurately in regular camera videos, such as those used to record standard tasks and those networked for civil surveillance. However, it remains unclear how they perform on the challenging wide area sequences in which we are interested. To this end, we carefully design a tracking evaluation experiment in this paper to test the applicability of visual trackers for wide area motion video. We first select sequences from the CLIF dataset and prepare the ground-truth target locations by manual annotation. Then, we tune several existing visual trackers to maximize their performance on these sequences. Note that the parameter tuning is required, since most of the methods by default only deal with real time sequences with small camera motion. Finally, these trackers are quantitatively evaluated and the results are analyzed. The tracking methods used in our experiments include the Mean Shift (MS) tracker [7], the Multiple Instance Learning (MIL) tracker [3], the intensity histogram based Particle Filter (HPF) [23], covariance tracker [25], and the Bounded Particle Resampling L1 (L1-BPR) tracker [20].

To focus on effects caused by low frame rates, we conduct a second group of experiments in complimentary to the above one. In the experiments, we first "calibrate" each video by aligning consecutive frames through registration, which uses scale invariant feature transform (SIFT) [16] and RANdom SAmple Consensus (RANSAC) [10] for warping estimation. Then we run the above tuned tracking algorithms on the aligned sequences.

There are mainly two observations from our experiments. First, visual trackers that work well for regular video sequences have trouble when applied to the extremely low frame rate videos. Second, the registration step does improve the performance. However, there is large room for improvement to validate visual trackers against robustness performance requirements.

In summary, our contribution in this paper lies in several aspects. First, we carefully evaluate existing visual trackers on wide area surveillance videos, which are taken under extremely low frame rate and large camera motion. The experiments provide an understanding of the challenges in the task. Second, our study show that SIFT+RANSAC effectively helps improve tracking performance. Finally, the evaluation framework, including the sequences and annotated ground truth, provides a useful benchmark for future studies.

The rest of the paper is organized as follows. In Section 2, we describe the tracking problem and the challenges faced by visual trackers. Next, the registration method is introduced in Section 3. Then, experimental results and analysis are provided in Section 4. Finally, we draw conclusions in Section 5.

2 Tracking in Wide Area Videos

In this section we first introduce the wide area surveillance video dataset used in our study. After that, we will investigate and illustrate the challenges for vehicle tracking tasks.

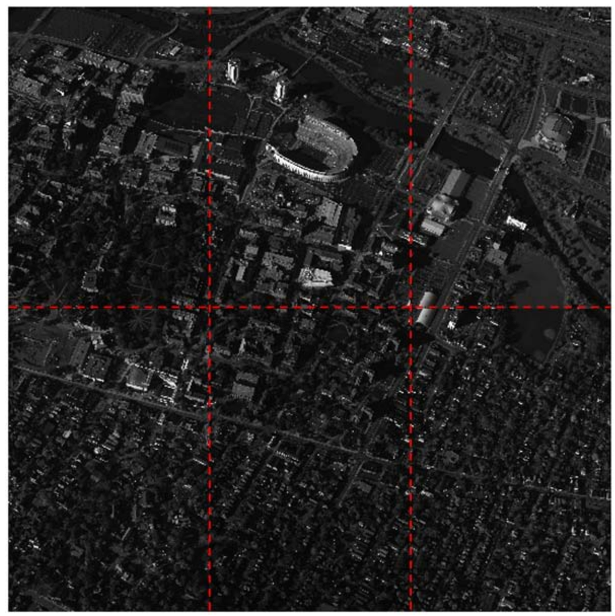
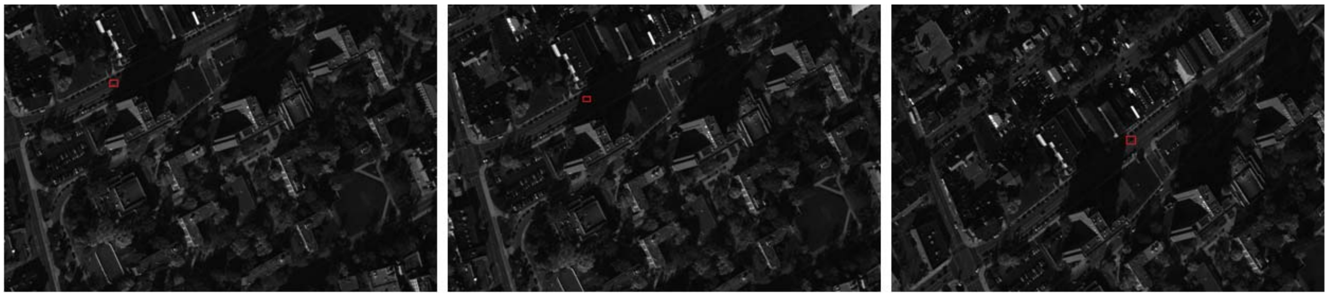


Figure 2. Example images from the AFRL CLIF dataset [1]. The dashed red line show the boundary of six images from different EO cameras.

2.1 Dataset

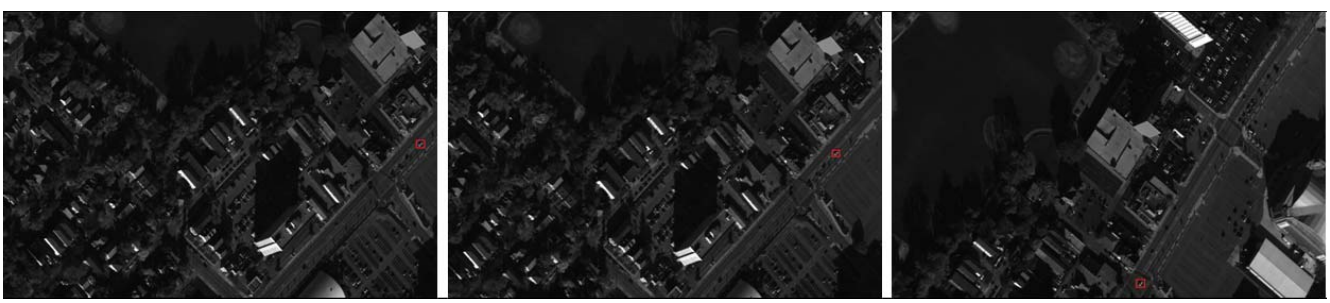
In our study we use the AFRL CLIF dataset (2007) [1] for experiments and illustration. The dataset includes three "layers" of data from multi-layered sources, including a high-flying asset, surrogate UAVs, and building mounted sensors. There are 96,702 (16,117 images x 6 cameras) raw files and 16,117 auxiliary text files. In this study we are interested in the video data collected from the high-flying asset. These videos are collected from a large format electro-optical (EO) platform. The scene in the dataset is a flyover of the Ohio State University (OSU) Campus using a large format monochromatic EO sensor, namely the Large Area Image Recorder (LAIR) sensor. The sensor is comprised of six cameras, with each of them oriented to maximize coverage but still allow overlap between images to enable mosaicing. Figure 2 shows example images take at the same time frame.

The sequences in CLIF are collected at approximately two frames per second, which is much lower than those used in the study of traditional visual tracking scenarios [28]. The size of each of the six images is 4016 by 2672, which is much larger than normal surveillance videos, let along the mosaic of all six images. Figure 3 shows exemplar images with ground truth annotations.

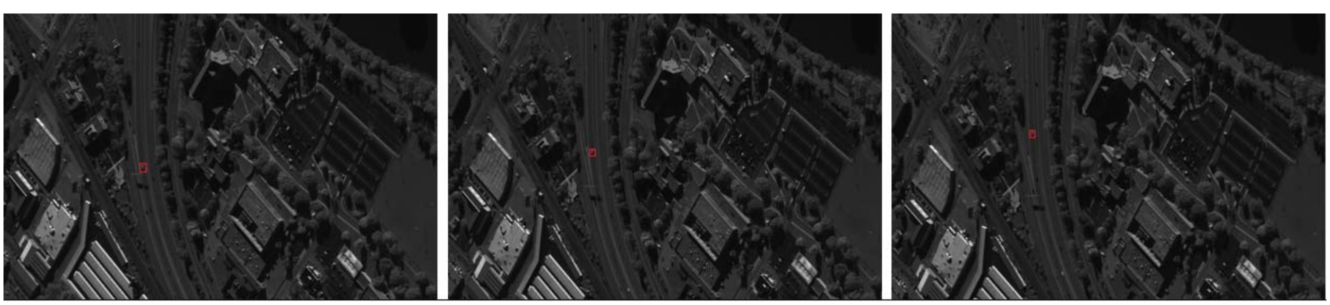


Large camera motion and shade occlusion. C0_3_0, frames 7, 9 and 17 from left to right.

Self pose change with camera motion. C4_1_0, frames 3, 13, and 20.



Vehicle motion against camera motion. C4_4_4, frames 7, 8 and 39 from left to right.



Fast vehicle motion. C5_1_4, frames 33, 34, and 35 from left to right.

Figure 3. Challenges of visual tracking in wide area motion imagery (examples the AFRL CLIF dataset [1]). Ground truth annotations are shown in red boxes.

2.2 Challenges

Tracking a vehicle through a wide area video is not an easy task due to many challenges. First of all, there are two fundamental reasons for these challenges: the extremely low frame rate and the drastic camera motion. The extremely low frame rate often causes a large target movement across frames, which in turn increases seriously the uncertainty in the prediction of target location. The drastic camera motion further exacerbates the problem. Furthermore, the two factors can cause discontinuity in target appearance change as well.

The general list of challenges with using the large format imagery is:

1. Large camera motion,
2. Low frame rate,
3. Shade (tree, building) and illumination,
4. Pose change of the target (car): up-right rectangle cannot model the car appropriately,
5. Adjacent clutter: similar objects (cars),
6. Background clutter,
7. Low contrast between the car and the background
8. Occlusion,
9. Tracking between image boundaries,
10. High resolution image format may request computational efficiency.

From this list, we see that using only the overhead imagery would be more difficult than stereo fusion of the imagery with ground mounted cameras to cue additional information for track validation. However, registering the ground information with the overhead information introduces additional requirements in communication and registration. We seek evaluation methods of using only the overhead imagery.

Figure 3 shows example frames from the CLIF dataset where the challenges are illustrated. From the examples we can see that these challenges often simultaneously exist, which brings extra difficulties to traditional visual tracking algorithms.

2.3 Testing Sequences and Ground truth

For quantitatively evaluating the tracking performance, we first selected 14 sequences, each of them containing at least one moving vehicles. The choice of these sequences is to cover as much as possible the above mentioned challenges.

Once the sequence set is ready, the target location in each frame of each sequence is manually annotated. The annotation is in the form of bounding boxes. Table 1 summarizes the properties of the sequences and corresponding ground truth.

Note that due to the rapid camera motion and low frame rate, in most of the sequences the target existing only in several dozen of frames. Nevertheless, this is fine for our evaluation purpose. In fact, as shown in our experiments, for most sequences, none of the tested trackers can successfully track through all frames.

Table 1 Summary of selected sequences from CLIF

Name	Start #	End #	Notes
C0_3_0	1	50	Shade occlusion
C1_2_0	24	50	Shade occlusion
C1_4_0	1	50	Shade occlusion
C1_4_6	1	50	Shade occlusion
C2_4_1	1	50	Shade occlusion
C3_3_4	10	36	Shade occlusion
C4_1_0	3	20	Left turn
C4_3_0	31	50	Shade occlusion
C4_4_1	21	50	Low contrast
C4_4_4	27	39	Motion against the camera motion
C5_1_4	26	49	Fast vehicle motion
C5_2_0	2	50	Fast vehicle motion
C5_3_7	23	49	Bus
C5_4_1	1	24	Low contrast

3 Registration

To further understanding the effects of the low frame rate on visual tracking, it is desired to compensate the camera motion through background registration. For this purpose, we use the SIFT+RANSAC for registration between continuous frames.

The *Scale Invariant Feature Transform* (SIFT), originally developed by David Lowe [16] at the University of British Columbia, uses SIFT features with an implementation of the RANSAC algorithm and a linear least squares fit. The method is used for frame-to-frame registration. The SIFT descriptors rely on an orientation assignment given to the detector, so the detector and the descriptor are not easily separable.

SIFT looks for local extrema in the scale-space which consists of several Difference of Gaussian (DoG) images. Extrema that are along edges or in low contrast regions are eliminated because they are considered to be unstable. The algorithm then convolves the area immediately around the keypoint with a Gaussian kernel and uses the result to calculate the local image gradients. The image gradients are then used to assign an orientation to the keypoint.

The descriptor is then computed as a 4×4 array of histograms. Each histogram contains 8 bins according to the original implementation of SIFT. The key to the descriptor is that it is calculated relative to the orientation of the keypoint by first resampling the area around the keypoint so that the orientation of the keypoint will be in one of the cardinal directions. By computing the descriptor relative to the orientation of the keypoint, rotational invariance is achieved. The descriptor is computed by convolving the area around the keypoint with a Gaussian kernel with sigma equal to 1.5 times that of the keypoint sigma. The image gradients at each pixel are calculated and the magnitude is added to the correct bin in the histogram, based on the direction of the image gradient.

The SIFT registration approach then uses RANSAC to find the optimal affine transformation between the image pairs based on the matching SIFT descriptors. The best transformation found by RANSAC is then used to select the points that are closest to each other after the transformation is applied. These points are then used to generate the final transformation between the two images by using a linear least squares fit.

Note that there are other registration methods tested on the CLIF dataset [18] and SIFT does not always achieves the best performance. However, in this study we focus on visual tracking and the SIFT+RANSAC method performances well enough for our task.

4 Evaluation

The evaluation of the visual trackers applied to the wide area imagery is subject to the registration and the metrics of evaluation.

4.1 Methods

Five state-of-the-art visual trackers are included in our evaluation. When applying these methods in our experiments, we tune the tracking parameters to maximize the performance, since these methods by default deals with standard videos. For example, the search ranges and number of candidates in these methods are enlarged to compensate to the large changes in target positions.

The five visual trackers are listed below:

- The *Mean Shift (MS) tracker* [7], which models tracking by an efficient mode seeking process;
- The *Multiple Instance Learning* (MIL) *tracker* [3], which models target inference through multiple instance learning;
- The *intensity histogram based Particle Filter* (HPF) [23], which a classical sequence Bayesian inference tracker using HSV color space;
- The *covariance based Particle Filter* (CPF) [25], which use the covariance information [24] to fuse different features for robust visual tracking;
- The *Bounded Particle Resampling L1* (L1-BPR) tracker [20], which efficiently brings sparsity into target representation to achieve robustness against appearance contamination [19].

4.2 Results

We conducted the above five visual trackers to the sequences described in Section 2.3. Each tracker is tested in two cases, one with background registration and the other without.

Figure 4 shows some typical results from the experiments. For quantitative study, we use the *missing frame rate* (MFR) to measure the tracking error, which is defined as

$$\text{MFR} = \frac{\text{number of missing frames}}{\text{number of total frames}} \qquad (1)$$

where a frame is treated as "missing" if the tracking result is overlapped with the ground truth by less than 50%.

From the results, we have several observations: First, both qualitative and quantitative results confirm that tracking through wide area motion imagery is a non-trivial problem and many issues remain open. Second, registration in general helps the tracking process and therefore is worth further investigating and coupling to the trackers.

Table 2. Quantitative evaluation in terms of missing frame rate (the smaller the better). The notation "+R" means that the tracker is combined with the cross frame registration described in Section 3.

Tracking without registration						**Tracking with registration**					
Seq.	MIL	MS	CPF	HPF	L1-BPR	Seq.	MIL+R	MS+R	CPF +R	HPF+R	L1-BPR+R
C0_3_0	**0.8600**	0.9800	0.9400	0.9800	0.9000	C0_3_0	0.8600	0.9800	0.9400	0.9800	**0.7600**
C1_2_0	**0.8519**	0.9630	0.9259	0.9259	0.9630	C1_2_0	0.8519	0.9630	0.9630	0.9630	**0.6296**
C1_4_0	**0.6800**	0.9800	0.8200	0.8600	0.7800	C1_4_0	0.6800	0.7800	0.7400	0.7600	**0.6200**
C1_4_6	0.9400	0.9800	0.9600	0.9200	**0.7800**	C1_4_6	**0.3600**	0.9400	0.8000	0.8800	**0.3600**
C2_4_1	**0.9800**	**0.9800**	**0.9800**	**0.9800**	**0.9800**	C2_4_1	**0.9000**	0.9800	0.9800	0.9800	0.9200
C3_3_4	**0.9630**	**0.9630**	**0.9630**	**0.9630**	**0.9630**	C3_3_4	0.9630	0.9630	0.9630	0.9630	**0.7037**
C4_1_0	**0.4444**	0.9444	0.9444	0.9444	0.9444	C4_1_0	**0.3889**	0.8889	0.8333	0.8889	**0.3889**
C4_3_0	**0.6500**	0.9500	0.9500	0.8500	0.7500	C4_3_0	**0.6500**	0.9500	0.9500	0.8000	0.7500
C4_4_1	**0.0333**	0.9667	0.9333	0.9333	0.9000	C4_4_1	0.5333	0.9667	0.9000	0.9000	**0.0333**
C4_4_4	**0**	0.9231	0.9231	0.9231	0.9231	C4_4_4	**0**	0.9231	0.3846	0.3077	**0**
C5_1_4	**0.7917**	0.9583	0.9167	0.9167	0.9167	C5_1_4	**0.6667**	0.9583	0.8750	0.8333	**0.6667**
C5_2_0	0.9592	0.9796	**0.8367**	0.9796	**0.8367**	C5_2_0	**0.9184**	0.9796	0.9592	0.9796	0.9796
C5_3_7	**0.0370**	0.9630	0.9259	0.9259	0.7407	C5_3_7	**0**	0.9630	0.1481	**0**	**0**
C5_4_1	**0.6190**	0.9524	0.8571	0.9048	0.8095	C5_4_1	**0**	0.9524	0.8095	0.9048	0.9583
All	**0.7105**	0.9693	0.9167	0.9342	0.8684	All	0.6272	0.9408	0.8333	0.8377	**0.6118**

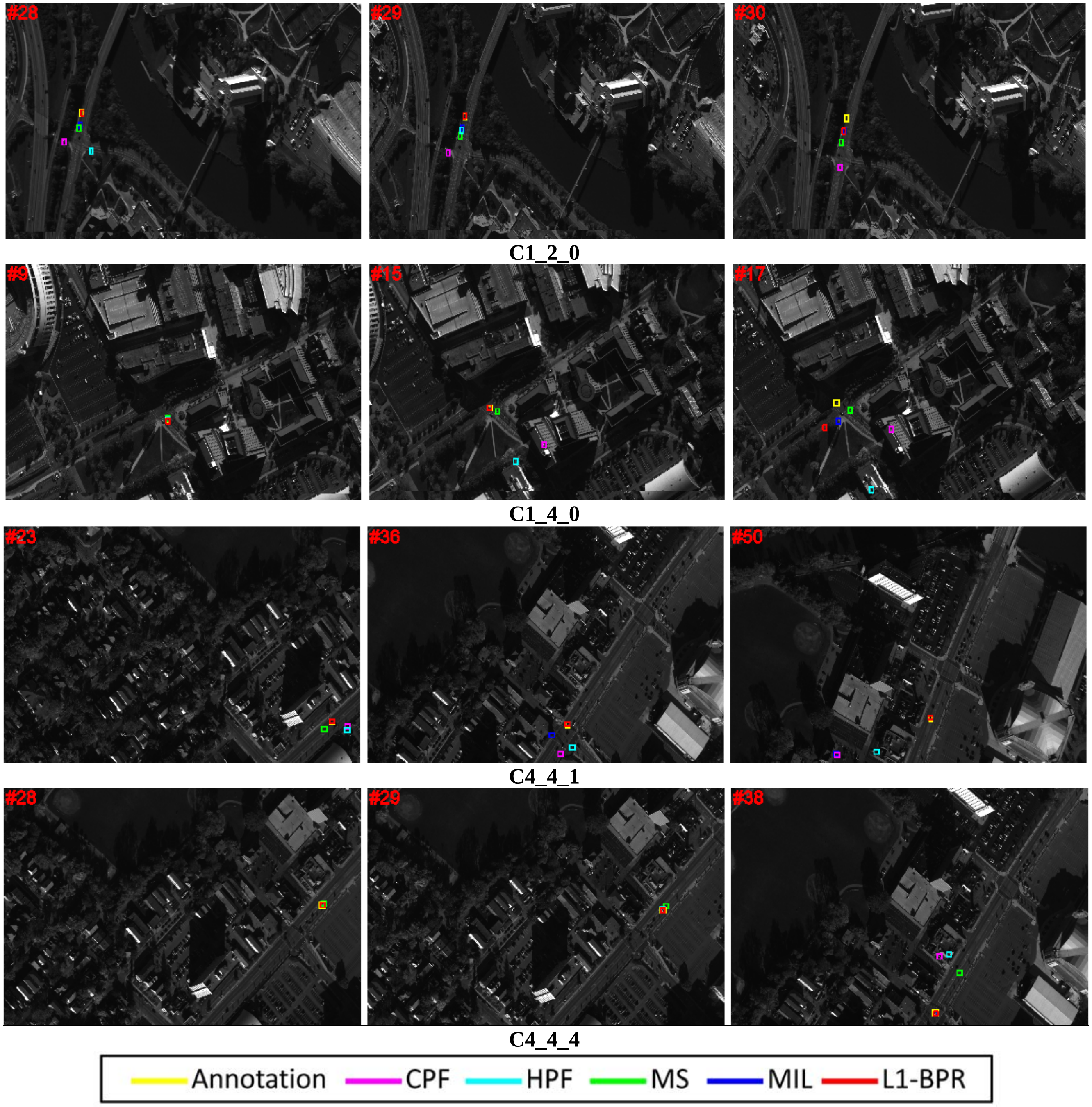


Figure 4. Example tracking results.

The best performance, is achieved by the MIL tracker without registration and by the L1-BPR with registration, while the worst by MS tracker. This can be well explained by the huge target displacement (in images) due to the extremely low frame rate and large camera motion. Intuitively, the search range of MIL can be easily enlarged to address the issue to some degree; while the MS tracker heavily depends on continuous target motion (in images) for mode seeking. On the other hand, when the registration is applied to alleviate the challenge, the L1-BPR method catches up and outperforms other methods. Other particle filter based methods work between the two extremes by adjusting the particle sample parameters. This phenomenon suggests that enlarging search range is a possible solution, however, such solution may suffer from the large computational request, especially considering the high resolution images used in our task. On the other hand, it also suggests that effective motion estimation may help largely reduce the search range, which is one direction we will investigate in the future.

5 Conclusion

In this paper we evaluated the performance of existing visual trackers on the wide area motion imagery (WAMI) scenario, which is very challenging due to many factors such as the extremely low frame rate and the large camera motion. We tested five state-of-the-art visual trackers in carefully selected and annotated sequences from the CLIF dataset, for the task of the vehicle tracking. The experimental results confirm our conjecture that traditional visual trackers meet serious problems in the wide area surveillance videos. In addition, the results show that background registration helps enhance the tracking performances.

We expect the study to motivate future research to addresses the issues we observed in the task. In particular, we are interested in modeling background information in the tracking, efficient joint registration and tracking, and robust trajectory estimation for drastic target motion.

6 Acknowledgment

This work is supported in part by NSF Grant IIS-1049032. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

References

[1] AFRL CLIF 2007 dataset over Ohio State University, https://www.sdms.afrl.af.mil/index.php?collection=clif2007

[2] S. Avidan. "Ensemble Tracking," *Proc. of the Computer Vision and Pattern Recognition* (CVPR), 2005.

[3] B. Babenko, M. Yang, and S. Belongie, "Visual tracking with online multiple instance learning," *Proc. of the Computer Vision and Pattern Recognition* (CVPR), 2009.

[4] M. J. Black and A. D. Jepson. "Eigentracking: Robust matching and tracking of articulated objects using a view-based representation," *International Journal of Computer Vision*, 26:63-84, 1998.

[5] E. Blasch and L. Hong, "Simultaneous Identification and Track Fusion," *IEEE Conf. on Dec. Control*, Orlando, FL, Dec 1998.

[6] E. Blasch and B. Kahler, "Multiresolution EO/IR Target Track and Identification", *ISIF Proc. Fusion05*, 2005.

[7] D. Chen and J. Yang. "Robust Object Tracking Via Online Dynamic Spatial Bias Appearance Models." *IEEE Trans. Pattern Anal. Mach. Intell.* 29(12): 2157-2169, 2007.

[8] D. Comaniciu, V. Ramesh, and P. Meer. "Kernel-based object tracking," *IEEE Trans on Pattern Anal. and Mach. Intell.*, 25:564-577, 2003.

[9] A. Doucet, N. de Freitas, and N. Gordon. *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, 2001, New York.

[10] M. A. Fischler and R. C. Bolles. "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography". *Comm. of the ACM*, 24: 381–395, 1981.

[11] G. D. Hager and P. N. Belhumeur. "Efficient region tracking with parametric models of geometry and illumination," *IEEE Trans on Pattern Anal. and Mach. Intell.*, 20:1025-1039, 1998.

[12] J. Ho, K.-C. Lee, M.-H. Yang, and D. Kriegman. "Visual tracking using learned subspaces," *in Proceedings of the Computer Vision and Pattern Recognition* (CVPR), 782-789, 2004.

[13] M. Isard and A. Blake. "Condensation - conditional density propagation for visual tracking," *International Journal of Computer Vision*, 29:5-28, 1998.

[14] B. Kahlerand E. Blasch, "Sensor Management Fusion Using Operating Conditions", *IEEE Aerospace & Electronics Conference (NAECON)*, 2009 National, 223--230, 2009

[15] E. K. Kao, M. P. Daggett, and M. B. Hurley. "An information theoretic approach for tracker performance evaluation," *International Conference on Computer Vision*, 1523-1529, 2009.

[16] H. Ling, L. Bai, E. Blasch, and X. Mei, "Robust Infrared Vehicle Tracking across Target Pose Change using L1 Regularization", *ISIF Proc. Fusion10*, 2010.

[17] D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, 60(2):91-110, 2004.

[18] O. Mendoza-Schrock, J. A. Patrick, and E. P. Blasch, "Video image registration evaluation for a layered sensing environment,", *IEEE Aerospace & Electronics Conference (NAECON)*, 2009 National, 223--230, 2009

[19] X. Mei and H. Ling. "Robust Visual Tracking using l_1 Minimization," *Proceedings of the International Conference on Computer Vision* (ICCV), 2009.

[20] X. Mei, H. Ling, Y. Wu, E. Blasch, and L. Bai. "Minimum Error Bounded Efficient L_1 Tracker with Occlusion Detection," *Proc. of the Computer Vision and Pattern Recognition* (CVPR), 2011.

[21] P. Minviell, A. D. Marrs, S. Maskell, and A. Doucet, "Joint Target Tracking and Identification – Part I:

Sequential Monte Carlo Model-Based Approaches, *ISIF Proc. Fusion05*, 2005.

[22] P. Minviell, A. D. Marrs, S. Maskell, and A. Doucet, “Joint Target Tracking and Identification – Part II: Shape Video Computing, *ISIF Proc. Fusion05*, 2005.

[23] P. Perez, C. Hue, J. Vermaak, and M. Gangnet. "Color-Based Probabilistic Tracking", in *Proceedings of the European Conference on Computer Vision* (ECCV), 2002.

[24] F. Porikli, O. Tuzel, and P. Meer. ``Covariance tracking using model update based on Lie algebra'', in *in Proceedings of the Computer Vision and Pattern Recognition* (CVPR), 2006.

[25] Y. Wu, B. Wu, J. Liu, and H. Q. Lu, "Probabilistic Tracking on Riemannian Manifolds," in *Proceedings of the IEEE International Conference on Pattern Recognition (ICPR), 2008.*

[26] C. Yang and E. Blasch, “Pose Angular-Aiding for Maneuvering Target Tracking,” *ISIF Proc. Fusion05*, 2005.

[27] C. Yang and E. Blasch. “Kalman Filtering with Nonlinear State Constraints.” *ISIF Proc. Fusion 07, 2007*.

[28] A. Yilmaz, O. Javed, and M. Shah. “Object tracking: A survey.” *ACM Comput. Survey*, 38(4), 2006.

[29] S. K. Zhou, R. Chellappa, and B. Moghaddam. “Visual tracking and recognition using appearance-adaptive models in particle filters,” *IEEE Trans. Image Processing*, 11:1491-1506, 2004.