Comparison of Infrared and Visible Imagery for Object Tracking: Toward Trackers with Superior IR Performance

Erhan Gundogdu, Huseyin Ozkan, H. Seckin Demir,
Hamza Ergezer, Erdem Akagündüz, S. Kubilay Pakin
Electro-Optics System Design Department, MGEO, ASELSAN Inc.
Ankara, Turkey
{egundogdu,huozkan,hsdemir,hergezer,eakagunduz,kpakin}@aselsan.com.tr

Abstract

The subject of this paper is the visual object tracking in infrared (IR) videos. Our contribution is twofold. First, the performance behaviour of the state-of-the-art trackers is investigated via a comparative study using IR-visible band video conjugates, i.e., video pairs captured observing the same scene simultaneously, to identify the IR specific challenges. Second, we propose a novel ensemble based tracking method that is tuned to IR data. The proposed algorithm sequentially constructs and maintains a dynamical ensemble of simple correlators and produces tracking decisions by switching among the ensemble correlators depending on the target appearance in a computationally highly efficient manner. We empirically show that our algorithm significantly outperforms the state-of-the-art trackers in our extensive set of experiments with IR imagery.

1. Introduction

Until recently, the cost of infrared (IR) sensors has been quite high such that they were mainly utilized in defense applications. However, due to the recent declining trend in the cost of IR sensors, system solutions for surveillance applications started to include IR cameras to their sensor suites. This recent ubiquity of IR cameras provided another type of data for video processing systems, acquired in an alternative spectrum. However, since the video processing systems have generally been researched for only visible band data, efficient exploitations of IR cameras as a valuable new tool of passive light/night imaging has remained largely unexplored. To that end, we study one of the major components in systems that utilize IR cameras: visual object tracking. Most object tracking methods [23], regardless of their approach, attempt to increase the performance of the tracker by addressing the issues such as occlusion, and fast changes in appearance and illumination. While some of these issues, such as occlusion, is common regardless of the spectrum in which the camera is capable of imaging, others, such as illumination, affect the images in rather different fashions. Thus, the optimal methods for videos acquired in one spectrum might turn out to be suboptimal for the others.

In this paper, we investigate the performance of the state-of-the-art object trackers and identify the IR specific challenges. There is a growing interest on comparing the characteristics of different imaging techniques in both the application level, (such as recognition [20], registration [8, 13, 22]) as well as the signal level [16]. For instance, a direct application of image processing techniques, which are specifically designed based on the visible spectrum, to the IR data is reported in [13, 14] to be non-optimal and results in performance losses. We emphasize that the presented study is the first to conduct such a comparative analysis in terms of the state-of-the-art trackers that are explicitly run on IR-visible video conjugates, i.e., video pairs captured through the same exact scene simultaneously. Based on the presented comparative study, we also propose a novel tracking method geared towards superior IR performance.

1.1. Related Work and the Comparison Bases

Visual object tracking is an extensively studied computer vision problem, however, mostly with visible imagery. A comprehensive benchmark study is reported in [23] for the-state-of-the-art as well as the most recently proposed tracking techniques. In feature based discriminative approaches, the problem is treated as a binary classification problem, cf. the pioneer studies [1, 7]. In this approach, positive and negative samples are bagged to discriminate the target from background by using descriptors such as Haar-like features [2, 9]. An example is the MILTrack algorithm [2], which combines weak classifiers to develop a strong one by exploiting the ideas from the multiple instance learning framework. In [25], a feature selector is further incorporated into the MILTrack method to obtain relatively more efficient implementations. FCT [24] uses the random projection idea of compressive sensing, where the track-by-

classification approach is applied after projecting the target image to a compressed domain. On the other hand, STRUCK [9] utilizes a kernelized structured output support vector machine (SVM). This method essentially incorporates the large margin theory into the tracking framework at the cost of a high computational load. As opposed to the standard SVM for binary classification, it uses continuous label information as in the case of regression and jointly kernelizes the label and the regressor, i.e., features. TLD [11] is also a discriminative approach, but it differs from other learning based approaches by its detection block. In Context tracker [6], context information is exploited using the regions called distracters and supporters. Distracters are regions whose target appearance are similar to target, whereas supporters are local keypoints around the target and have similar motion pattern.

In generative approaches, a model is built for the target appearance and, best candidate is searched for each frame via a search mechanism. IVT [21] is a method which learns an appearance based model using eigenbasis representation within a particle filter framework. The main contribution of this work is the proposed incremental PCA algorithm in which, as the appearance changes, the eigenbasis vectors are updated incrementally. In [15], multiple appearance and motion models are combined to model the status of object. Sparse principal component analysis is utilized to model the different realizations of the object.

Despite the recent popularity of discriminative approaches, correlation based approaches are also significant due to their high computational efficiency. MOSSE [4] is an adaptive correlation based tracking method where the convolution theorem is exploited and an optimum filter is designed using a set of target image samples. This algorithm is the most efficient algorithm among the aforementioned methods in terms of computational burden. In [10], a hybrid approach is proposed. Random samples in track-by-classification algorithms are replaced with dense samples designed to induce a circulant structure.

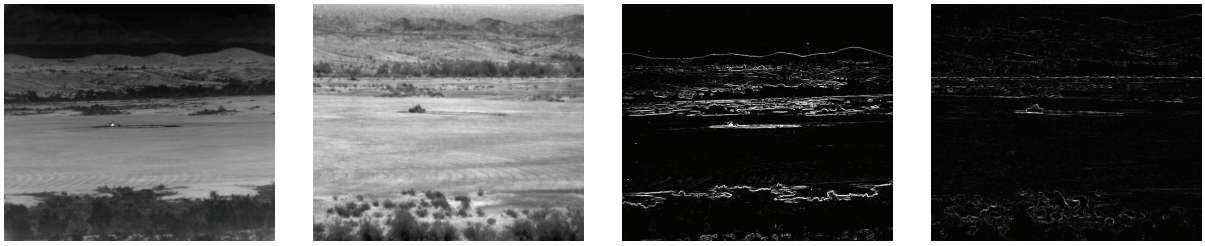
More recently, target object is sparsely represented with trivial and target templates in [19], where the sparse representation is obtained by solving L1-norm related minimization problem. In [17], a two-stage optimization is used to reconstruct candidate sample from templates. In [3], the method [19] is improved by introducing a new L1-norm related minimization model and utilizing proximal algorithms to efficiently minimize introduced cost function. Most sparse representation based approaches [3, 17, 19] exploit particle filtering to propagate sample distributions over time causing a significant growth in the computation time.

Depending on the system design, tracking algorithms can be run either on the imaging system, especially if the bandwidth is limited for data transfer, or on a central computer. The former case is especially important for IR videos. Placing the IR camera in a gimbal so that the tracked object is kept at the center of the acquired image is an important and common application in defense systems. To achieve this, trackers has to run real-time on the embedded platform whose computational and memory resources are somewhat limited. In such cases, the ability to run real-time at high frame rates becomes a crucial property for the tracker. Thus, our main concerns about tracking with IR imagery are not only the tracking performance but also the computational efficiency. For this reason, template based correlator trackers [4,10] are included in our comparative study due to their impressively efficient implementations. In addition, the feature based discriminative approaches [2, 24, 25] provide a comparison basis to our study, since the extraction of the Haar-like features in these approaches allows the use of efficient integral images. We omit the complex features such as HOG [5] and SIFT [18] since the extraction of them hardly satisfies our processing efficiency requirements. Nevertheless, we also consider the method [9], which also uses Haar-like features but in a more sophisticated classification framework via the max margin theory. Lastly, two representatives [3, 21] from other template based methods that are not correlators but generative are chosen as another comparison basis.

1.2. Toward Trackers with Superior IR Performance

In the presented comparative study, we perform experiments with the state-of-the-art tracking algorithms [2–4, 9, 10, 21, 24, 25] on IR as well as visible spectrum video pairs that are simultaneously acquired observing the same scene. Our aim is to investigate the performance behaviour of each compared method when the imaging spectrum is switched from visible to IR and identify the IR specific challenges. Afterwards, based on our findings, we develop a tracker to satisfy both performance and computational complexity requirements in IR data.

In our experiments, we observe a dramatic performance loss associated with feature based discriminative tracking approaches, whereas template based simpler correlators turned out to be less sensitive to the imaging spectrum. Several properties of the IR images are the reason for this behaviour. IR images, especially as the imaging wavelength gets longer, are dominated by the blackbody radiation of the observed objects, rather than the reflective component (it is the opposite for images in visible spectrum). The emissivity and the temperature of the objects are the factors that determine the amount of the blackbody radiation and these factors, in general, change smoothly across the objects. Thus, IR images are less textured compared to their visible counterparts (An object with a strong heat source, such as an engine, might be seen as an exception. However, even in that case, its bright spots that correspond to the heat sources



(a) IR band image (b) Visible band image (c) IR band edge map (d) Visible band edge map

Figure 1. Edge strength for an IR/visible band image pair.

make it easy to correlate in between frames). Moreover, the edge responses in an IR image are typically more noisy and hence background clutter strongly conceals the target features that are already not very strong in the IR image. In Figure 1, it is illustrated that IR generates edges both on the foreground and the background (i.e. contamination by clutter), whereas in the visible image, the contours of the target can be easily segmented. On such a scene, feature based tracking methods are adversely affected in the IR spectrum. To overcome this issue of scarcity of discriminative features, overly dense HOG features are used in another context of object recognition [12] with IR images. Such densely sampled HOG features might also be a solution for object tracking, but it would violate the real-time processing requirements of the security and defense applications. Instead, the template based correlator trackers utilize the full image representations -in line with the over sampling idea- and also allow fast FFT algorithms in generating track decisions that are robust to changes in imaging spectrum. Nevertheless, template based correlators have limited modelling power compared to the sophisticated classification schemes of feature based discriminative approaches. For instance, while the correlation filter in [4] utilizes only one projection onto the template, the structured SVM method in [9] produces a kernel expansion enjoying projections as many as desired. Therefore, for object tracking in IR videos, either features specifically tuned to IR signal characteristics are to be efficiently designed or the modelling power of simple correlators is to be boosted with controllable computational complexity.

We opt to concentrate on the simple template based correlators and propose a novel ensemble method for tracking in IR. The proposed method is an adaptive switching mechanism in an ensemble of base correlators, i.e., base trackers, each of which utilizes a certain target template and registers a certain target appearance. At each time in the course of the tracking, our algorithm chooses the right base tracker depending on the active appearance mode of the target, updates only the chosen base tracker and produces a tracking decision by mainly using that of the trained tracker. As a result, the proposed framework is not only as computationally efficient as the individual base trackers but significantly boosts the modelling power of one single correlator while exploiting its representational superiority over feature based approaches. In our experiments, we empirically show that our algorithm significantly outperforms the state-of-the-art trackers in IR videos.

We explain the details of the proposed method in the following Section 2. Then, we present our comparative study and experiments in Section 3. The paper concludes with final remarks in Section 4.

Algorithm 1 Proposed TBOOST Algorithm

Input: Video: v_t, Budget: M, Number of frames: N

Initialize tracker and template sets $\mathbf{T} \leftarrow \emptyset, \mathbf{D} \leftarrow \emptyset$
Get the user input x_1
Add first tracker T_1 and the template D_1:
$\mathbf{T} \leftarrow \{T_1\}, \mathbf{D} \leftarrow \{D_1\}$
for $t = 2 \rightarrow N$ **do**
    Obtain the image patch x_t
    Map x_t onto $\mathbf{D}$ using 1
    $i = \underset{j}{\operatorname{argmax}}\ \mathbf{a_T}(j)$
    Calculate appearance change score $s = f(D_i, x_t)$
    if s is sufficiently large **then**
        Update tracker T_i
    else if $|\mathbf{T}| < M$ **then**
        $l \leftarrow |T| + 1$
        Add new tracker and the template
        $\mathbf{T} \leftarrow \mathbf{T} \cup \{T_l\}, \mathbf{D} \leftarrow \mathbf{D} \cup \{D_l\}$
    else
        $i = \underset{j}{\operatorname{argmin}}\ \mathbf{a_T}(j)$
        and replace T_i by a new tracker and template
    end if
    Perform tracking using the combined tracker
end for

Return: Target Locations

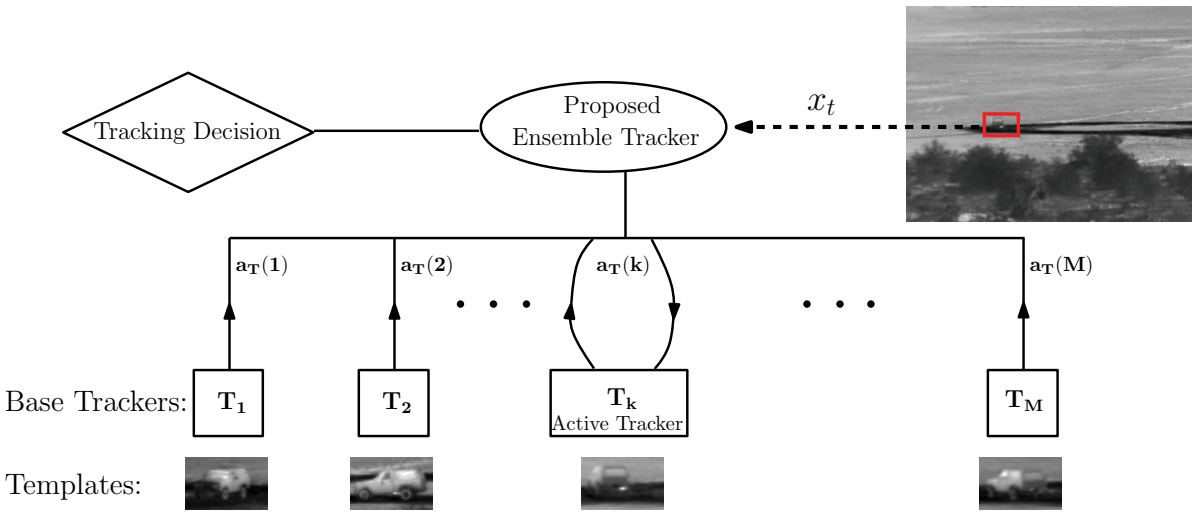


Figure 2. Illustration of the proposed TBOOST algorithm. TBOOST sequentially constructs and maintains a dynamical ensemble of base trackers, each of which is specialized on a different target appearance, such that a complete history of tracking experiences is registered in the course of the video stream. The proposed algorithm, TBOOST, implements a non-smooth, highly adaptive and time varying learning rate by continuously switching in the ensemble to choose the right base tracker. This switching mechanism is in accordance with the target appearance dynamics. At each frame t, only the active (chosen) base tracker is updated with only the informative part of the received image patch (the noise or occluded part is subtracted) and the final tracking decision by TBOOST is collectively based on the ensemble. See the text for details.

2. Proposed Method

A tracker is generally designed to adaptively learn the target appearance at usually a user specified learning rate. The specification of this adaptation rate must ideally be in accordance with the actual rate of the appearance change in the scene, which -however- cannot be foreseen and is not necessarily constant. Namely, the appearance change can realize at various rates depending on the scene dynamics and there exists no one optimal adaptation rate that a tracker can choose a priori. Therefore, a time varying adaptation must be incorporated and we implement such an approach using an ensemble method.

To this end, we propose an algorithm, named "TBOOST" presented in Algorithm 1, that adaptively switches in an ensemble of base trackers, where -at any time t- only one base tracker actively learns the target appearance. The proposed procedure sequentially constructs and maintains a dynamical ensemble of base trackers, each of which is specialized on a certain target appearance. At any time t, if the active base tracker is sensed to unable to effectively adapt to the target appearance, then the proposed algorithm switches to the most appropriate base tracker from the ensemble. If none of the ensemble trackers is found to be appropriate, a new base tracker is initialized and included in the ensemble. We also keep the ensemble size, i.e., budget size, fixed by using a certain add/removal strategy. We point out that the introduced switching mechanism effectively implements a non-smooth highly adaptive learning rate. Moreover, note that the proposed framework keeps an history of tracking experience during the video stream, and allows recalling an appearance model whenever it re-emerges instead of completely forgetting or discarding it.

We secondly emphasize that the proposed framework specially handles both the (additive Gaussian type of) unstructured or structured (such as occlusion) noise that might exist in the target appearance. Clearly, such a noise component is detrimental on the tracking performance and not only the switching among the base trackers but learning/updating the active base tracker mechanisms should ideally not be affected by this noise component. For this purpose, we exploit the recently proposed L1 projection technique [3]. Lastly, the final tracking decision is obtained by combining all of the base trackers in the ensemble and the computational complexity of the introduced method is not more than that of the individual base trackers. The proposed tracking algorithm is summarized in Algorithm 1 and illustrated in Fig. 2; and in the following we explain the details.

We assume a video stream v_t of gray scale frames.[1] Let the image patch at each frame cropped by the tracking decisions be denoted by x_t, where the first image patch x_1 is provided (as a box covering the target object) by the user. Given the budget size M, the first base tracker T_1 is initiated with the provided image patch x_1. For this base tracker T_1, we also register the image patch x_1 as the corre-

[1] Our algorithm can be straightforwardly extended to operate on standard videos of 3-channel RGB frames.

Figure 3. Samples from SENSIAC (first three columns) and OTCBVS (last column) datasets

sponding $D_1 = x_1$ appearance model. In the course of the tracking, we maintain a dynamical ensemble of base trackers $\mathbf{T} = \{T_i\}_{i=1}^{M}$ and a corresponding set $\mathbf{D} = \{D_i\}_{i=1}^{M}$ of appearance models. In general, when the frame v_t and hence the image patch x_t based on the previous track decision arrives, we use the L1 projection method [3] to decide which base tracker in the ensemble $\mathbf{T}$ to switch to and update. Namely, we obtain a sparse representation for the appearance x_t in terms of the ones in the template set $\mathbf{D}$ via

$$\underset{\boldsymbol{a}}{\operatorname{argmin}} \left\{ ||\boldsymbol{x} - \mathbf{K}\boldsymbol{a}||^2 + \lambda||\boldsymbol{a}||_1 + \mu||\boldsymbol{a_I}||_2 \right\}, s.t. \boldsymbol{a} \succeq 0, \tag{1}$$

where $\boldsymbol{a} = [\boldsymbol{a_T}, \boldsymbol{a_I}]^T$ and $K = [\mathbf{D}, \mathbf{I}]$. Note that with this L1 projection, we aim to lump the noise in the currently received appearance x_t such as an occlusion in the noise image $\mathbf{a_I}$ and obtain the sparse representation $x_t \simeq \mathbf{D}\boldsymbol{a_T}$. We then conclude that if the appearance model with the maximum weight in $\mathbf{a_T}$ is sufficiently similar to the x_t[2], then the template $D^* \in \mathbf{D}$ is the best representing one, i.e., let the appearance D^* corresponds to $\max \mathbf{a_T}$, and the corresponding base tracker $T^* \in \mathbf{T}$ must be active and updated with the received observation x_t. Otherwise, if the budget size has not yet been exceeded, a new base tracker is initialized and its template is set as the current observation x_t. If the budget is full and we do not have a sufficiently good representation, i.e., the appearance model with maximum weight in $\mathbf{a_T}$ is not sufficiently similar to the current observation x_t, then the base tracker corresponding to the lowest coefficient in $\boldsymbol{a}_T$ is replaced with a newly initiated base tracker.

In our framework, we do not restrict the choice of the base tracker, and any tracker can be used for this purpose. However, for its high performance and robustness on IR images as well as its computational efficiency, we use the MOSSE filters as our base tracker. In the following, we briefly summarize the MOSSE tracking approach and then present our final combined tracker.

We use the MOSSE filters [4] as our base trackers without loss of generality. In the tracking framework of [4], a set of training images $\{f_i\}$ of the target object is assumed to be provided beforehand[3]. Then, the goal is to find a correlator h such that the correlation between any f_i and the correlator h yields a relatively large response. To obtain such an optimal correlator h and the corresponding optimal filter response H (after the Fourier transform), the following optimization is performed:

$$H = \underset{H^*}{\operatorname{argmax}} \sum_i | F_i \odot H^* - G_i |^2, \tag{2}$$

where capital case notations indicates the Fourier counterparts and G_i is the desired response (details can be found in [4]). Therefore, since we use a basic MOSSE filter for each of our base tracker, we have a corresponding set of $\{H_i\}_{i=1:M}$ at any time t. Then we construct an average filter $H_{avg} = \Sigma_{i=1:M} \mathbf{a_T}(i) H_i$[4] and obtain the decision of the combined proposed ensemble tracker by locating the peak of $F^{-1}({H_{avg}}^* \odot X_t)$ (F^{-1} denotes the inverse FFT), which basically yields the location of the target object.

The computational complexity of the localization part of the algorithm is not more than that of the base tracking algorithm, i.e., MOSSE in this paper ($Plog(P)$ with P template dimension). Secondly, L1 projection is performed at each frame to find the representation coefficients of the base trackers. For this purpose, we use the optimization method used in [3] once at each frame. Therefore, our method has a computational complexity of MOSSE [4] in addition to complexity of the L1 projection per frame. Unlike the sparse coding based visual tracking methods fulfilling L1 norm minimization at each frame as many times

[2]By comparing the normalized cross correlation between them to an appropriate threshold

[3]This set of training images is typically obtained by applying random perturbations to the initial target patch provided by the user

[4]By the formulation in 1, $\mathbf{a_T}$ is non-negative, and here it is normalized to add up to 1 for our averaging purposes.

Table 1. Overall Tracking Performance Change from Visible to IR Sequences. The methods are listed in the descending order of the performance loss. The methods in the first four rows are feature based, the remaining ones are template based.

	AUC	*TM*	*Precision*
MIL	-63.91 %	-54.19 %	-48.79 %
ODFS	-46.52 %	-39.67 %	-34.00 %
FCT	-38.09 %	-38.09 %	-33.85 %
STRUCK	-28.89 %	-22.76 %	-24.77 %
L1APG	-13.76 %	-13.95 %	-3.31 %
TBOOST	**-8.82 %**	**-3.10 %**	**-10.95 %**
MOSSE	-1.67 %	-12.25 %	-2.69 %
CRC	15.66 %	-19.07 %	-5.32 %
IVT	16.51 %	11.10 %	27.74 %

as the number of particles [3, 17, 19], the proposed method requires only one minimization procedure. We experimentally observe that our technique operates at around 40 fps (with a budget size of 10) in MATLAB environment with a 3.2 GHz Intel processor.

3. Experiments

Two groups of experiments are carried out for the performance analysis, where we compare the proposed TBOOST algorithm with several state-of-the-art trackers: MILTrack [2], ODFS [25], FCT [24], STRUCK [9], L1APG [3], MOSSE [4], CRC [10] and IVT [21], cf. Section 1 for a discussion about these methods. We use publicly available source codes provided in [23]. In the first set of experiments, all of the compared methods are run on both the IR and visible band sequences (captured observing the same scene simultaneously) of SENSIAC and OTCBVS datasets [5]. In the consequent set of experiments, the effect of target sizes on the tracking performance is investigated using IR band sequences.

3.1. Performance metrics

In our performance analysis, we use two evaluation metrics, i.e., success and precision rates, that are used in [23].

One of these metrics is the success rate that indicates the percentage of frames, in which the overlap (in terms of the bounding boxes) between the ground truth and the tracking result is sufficiently high with respect to an appropriate threshold. A success rate plot demonstrates this percentage for varying thresholds in $[0, 1]$. To rank the methods based on their success rates, we use the area under curve (AUC) and track maintenance (TM) scores, which are derived from success plots. AUC refers to the total area under a success rate plot, whereas TM is the ability of a tracker to maintain a track, i.e., the success rate at the 0^+ threshold (TM is given as % percentage on the tables).

[5]Refer to {www.sensiac.org} and {www.vcipl.okstate.edu} for the datasets SENSIAC and OTCBVS, respectively.

Table 2. SENSIAC - Success Rate Comparison. Although STRUCK operates well in visible data, its performance degrades considerably in IR. On the other hand, our method TBOOST maintains its success when switched from visible to IR and also outperforms its competitors.

	IR		Visible	
	AUC	*TM*	*AUC*	*TM*
TBOOST	**0.327**	**78.73**	**0.360**	**81.47**
STRUCK	0.297	63.65	0.420	82.71
MOSSE	0.211	57.79	0.215	66.07
L1APG	0.202	47.50	0.235	55.39
FCT	0.178	44.20	0.289	71.88
IVT	0.127	35.00	0.109	31.58
ODFS	0.120	33.83	0.225	56.51
CRC	0.119	27.18	0.103	33.82
MIL	0.055	16.25	0.154	35.83

Table 3. SENSIAC - Precision Comparison

	IR	Visible
TBOOST	**66.85**	**74.74**
L1APG	58.14	60.09
STRUCK	57.50	76.27
MOSSE	51.78	53.18
FCT	45.20	66.21
IVT	38.76	30.09
ODFS	32.89	49.10
CRC	29.24	29.81
MIL	17.08	33.03

The other evaluation metric is the precision, which denotes the percentage of the frames, in which the standard Euclidean distance between estimated and actual target centers is sufficiently small with respect to an appropriate threshold. This evaluation metric demonstrates the localization accuracy of a method. To rank methods based on their precision, a distance threshold of 20 px is used.

3.2. Datasets

The SENSIAC dataset includes simultaneously captured visible and mid-wave IR sequences of certain scenes with various types of target objects of different target sizes such as walking pedestrians, pickup trucks, tanks and others. A ground truth that contains the target bounding boxes for each frame is also provided. In our experiments, we use 20 pairs of sequences, which contains considerable amount of background clutter in addition to several occlusion instances in case of walking pedestrians (Figure 3).

The OTCBVS dataset includes 4 indoor scenes of human-walking patterns (similarly, captured simultaneously with both visible band and mid-wave IR cameras). All images are rectified to register pixel correspondences between visible band and IR frames. This dataset is extremely challenging in the sense that the targets often occlude each other throughout the scenes (Figure 3). We manually gener-

Table 4. OTCBVS - Success Rate Comparison

	IR		Visible	
	AUC	*TM*	*AUC*	*TM*
STRUCK	0.126	29.73	0.065	22.07
TBOOST	**0.113**	**26.97**	**0.062**	**17.78**
FCT	0.092	26.50	0.053	12.40
L1APG	0.083	20.18	0.047	13.64
MOSSE	0.082	24.64	0.063	17.25
CRC	0.077	23.78	0.062	14.50
ODFS	0.076	24.05	0.051	11.98
MIL	0.072	19.78	0.058	18.63
IVT	0.065	17.43	0.050	11.15

Table 5. OTCBVS - Precision Comparison

	IR	Visible
TBOOST	**10.88**	**5.31**
STRUCK	8.45	3.77
L1APG	8.26	4.94
FCT	6.92	3.87
MOSSE	6.14	4.67
IVT	5.81	4.64
MIL	5.00	4.29
ODFS	4.19	4.20
CRC	4.32	5.68

ate a precise ground truth by annotating a bounding box for each person in the sequences.

3.3. Relative Performance between IR and Visible

We thoroughly discuss in Section 1 that the compared methods can be studied under two main families: feature based trackers and template based trackers. Since the dominant background clutter in IR band detrimentally affects the performance of feature based methods as a result of the suppressed foreground (target) features due to the IR characteristics (cf. Section 1), we observe a relatively large performance variations in case of feature based trackers in Table 1. Namely, template based trackers tend to preserve the visible band performance when the imaging technique changes to IR.

Next, we present our detailed comparisons separately for the two datasets.

3.3.1 Performance on SENSIAC Dataset

The overall performance results on SENSIAC dataset are depicted in Figure 4 (a - d) as well as Table 2 and Table 3. We observe that the performance of template based methods on IR sequences is relatively higher than that of feature based methods. Some template based methods such as MOSSE, TBOOST and CRC are limitedly affected by the radiation spectrum. On the contrary, feature based approaches such as STRUCK, FCT, ODFS and MILTrack experience dramatical performance losses up to 60 percent both in terms of success rate and precision. Namely, template based methods' ability to localize and maintain a track are considerably better in IR band (Table 2 and Table 3).

Table 6. Small vs Big Target. Our method TBOOST significantly outperforms its nearest competitor STRUCK especially for small targets.

	Small		Big	
	AUC	*TM*	*AUC*	*TM*
TBOOST	**0.241**	**65.88**	**0.349**	**81.94**
STRUCK	0.131	27.14	0.338	72.78
MOSSE	0.148	39.86	0.227	62.28
L1APG	0.160	48.00	0.213	47.37
FCT	0.182	30.02	0.177	47.74
IVT	0.027	10.41	0.152	41.14
ODFS	0.090	14.89	0.127	38.56
CRC	0.057	15.57	0.134	30.09
MIL	0.035	5.03	0.060	19.05

Most importantly, the proposed method TBOOST performs best in the IR band in terms of success rate and precision. Although it shows a slightly lower performance than STRUCK in visible band, its robustness to imaging technique makes TBOOST outperform its nearest competitor STRUCK in IR in terms of both metrics.

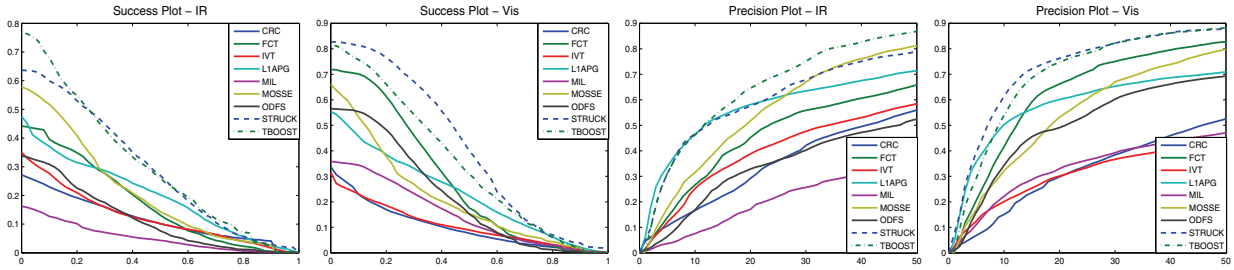
3.3.2 Performance on OTCBVS Dataset

The performance results on OTCBVS (Figure 4 (e - h), Table 4 and Table 5) show that the proposed method TBOOST performs better than all of the compared methods on IR sequences in terms of precision; and ranks in the second place in the success rate table. However, the performances of all methods decrease considerably on this dataset, since the OTCBVS is a remarkably challenging and difficult one due to its extreme amounts of occlusion. Thus, it is -in fact- difficult to reason about the performance changes of feature based trackers and draw concrete conclusions.

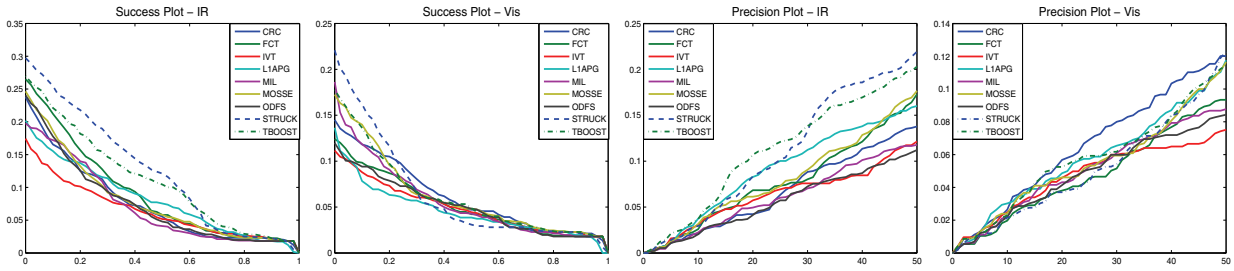
3.4. Target Size based Performance Analysis

To demonstrate the effect of target size on the tracking performance, another group of experiments are performed on IR sequences. The SENSIAC dataset is grouped into two groups; one with targets smaller than 100 pixels area, and the other with larger.

As the targets get smaller in size, the features on the target get less effective. Figure 5 shows the success and precision plots for these experiments and Table 6 shows AUC and TM scores. As targets get smaller and hence the discriminative features vanish, the proposed TBOOST algorithm outperforms the compared methods even more notably. Interestingly, STRUCK loses its second best position in overall performance, once again showing the negative correlation between the IR band and feature based tracking.

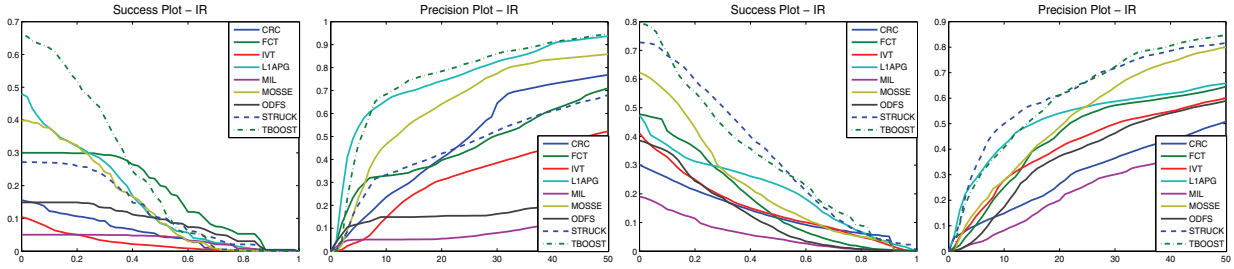


(a) SENSIAC-IR (b) SENSIAC-Vis (c) SENSIAC-IR (d) SENSIAC-Vis



(e) OTCBVS-IR (f) OTCBVS-Vis (g) OTCBVS-IR (h) OTCBVS-Vis

Figure 4. Overall Results



(a) smaller than 100 px-IR (b) smaller than 100 px-IR (c) bigger than 100 px-IR (d) bigger than 100 px-IR

Figure 5. Performance for bigger/smaller than 100 px.

4. Conclusion

We compare the infrared (IR) and visible imagery for object tracking via extensive experiments and propose a novel tracking algorithm that is tuned to IR data. Our development toward the proposed tracker with superior IR performance is based on the following observations: i) The typical security and defense applications of IR requires high computational efficiency with minimal memory usage. This signifies the simple template correlator based trackers which allows the efficient FFT algorithms. ii) The efficiently extractable features such as the Haar-like are faded away and cluttered by the IR cameras. Full image representations, which the simple template based correlators depend on, mitigate this issue to a significant degree. iii) Despite their impressive efficiency, the modelling power of these correlators are limited in the space of target appearances and need be boosted. To this end, the proposed algorithm adaptively and continuously switches in a specially designed ensemble of correlators depending on the observed target appearances in the course of the tracking. We show that via extensive set of experiments, our algorithm significantly outperforms the state-of-the-art techniques with IR imagery.

References

[1] S. Avidan. Ensemble tracking. In *CVPR*, 2005. 1

[2] B. Babenko, M.-H. Yang, and S. Belongie. Visual tracking with online multiple instance learning. In *CVPR*, pages 983–990, 2009. 1, 2, 6

[3] C. Bao, Y. Wu, H. Ling, and H. Ji. Real time robust l1 tracker using accelerated proximal gradient approach. In *CVPR*, pages 1830–1837. IEEE, 2012. 2, 4, 5, 6

[4] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui. Visual object tracking using adaptive correlation filters. In *CVPR*, pages 2544–2550. IEEE, 2010. 2, 3, 5, 6

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Computer Society Conference on CVPR 2005.*, volume 1, pages 886–893, 2005. 2

[6] T. B. Dinh, N. Vo, and G. Medioni. Context tracker: Exploring supporters and distracters in unconstrained environments. In *CVPR*, 2011. 2

[7] H. Grabner, M. Grabner, and H. Bischof. Real-time tracking via on-line boosting. In *BMVC*, 2006. 1

[8] J. Han, E. Pauwels, and P. de Zeeuw. Visible and infrared image registration employing line-based geometric analysis. In *Computational Intelligence for Multimedia Understanding*, pages 114–125. Springer, 2012. 1

[9] S. Hare, A. Saffari, and P. H. Torr. Struck: Structured output tracking with kernels. In *ICCV*, pages 263–270. IEEE, 2011. 1, 2, 3, 6

[10] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista. High-speed tracking with kernelized correlation filters. *IEEE Transactions on Image Processing*, 2014. 2, 6

[11] Z. Kalal, K. Mikolajczyk, and J. Matas. Tracking-learning-detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(7):1409–1422, 2012. 2

[12] M. Khan, G. Fan, D. R. Heisterkamp, and L. Yu. Automatic target recognition in infrared imagery using dense hog features and relevance grouping of vocabulary. In *IEEE Conference on CVPR Workshops*, June 2014. 3

[13] S. J. Krotosky and M. M. Trivedi. Mutual information based registration of multimodal stereo videos for person tracking. *Computer Vision and Image Understanding*, 106(2):270–287, 2007. 1

[14] S. Kumar, T. K. Marks, and M. Jones. Improving person tracking using an inexpensive thermal infrared sensor. In *CVPR PBVS Workshop*, June 2014. 1

[15] J. Kwon and K. M. Lee. Visual tracking decomposition. In *CVPR*, 2010. 2

[16] X. Li and N. Aouf. SIFT and SURF feature analysis in visible and infrared imaging for uavs. In *IEEE 11th International Conference on Cybernetic Intelligent Systems (CIS)*, pages 46–51. IEEE, 2012. 1

[17] B. Liu, L. Yang, J. Huang, P. Meer, L. Gong, , and C. Kulikowski. Robust and fast collaborative tracking with two stage sparse optimization. In *ECCV*, 2010. 2, 6

[18] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 2

[19] X. Mei and H. Ling. Robust visual tracking using l1 minimization. In *ICCV*, 2009. 2, 6

[20] A. Mian. Comparison of visible, thermal infra-red and range images for face recognition. In *Proceedings of the 3rd Pacific Rim Symposium on Advances in Image and Video Technology*, PSIVT '09, pages 807–816, Berlin, Heidelberg, 2008. 1

[21] D. A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang. Incremental learning for robust visual tracking. *International Journal of Computer Vision*, 77(1-3):125–141, 2008. 2, 6

[22] S. Sonn, G.-A. Bilodeau, and P. Galinier. Fast and accurate registration of visible and infrared videos. In *CVPR PBVS Workshop*, June 2013. 1

[23] Y. Wu, J. Lim, and M.-H. Yang. Online object tracking: A benchmark. In *CVPR*, 2013. 1, 6

[24] K. Zhang, L. Zhang, and M. Yang. Fast compressive tracking. *IEEE Transactions on Image Processing*, 2014. 1, 2, 6

[25] K. Zhang, L. Zhang, and M.-H. Yang. Real-time object tracking via online discriminative feature selection. *IEEE Transactions on Image Processing*, 22(12):4664–4677, 2013. 1, 2, 6