Vehicle Tracking in Wide Area Motion Imagery using KC-LoFT Multi-Feature Discriminative Modeling

Noor M. AL-Shakarji†‡, Filiz Bunyak*, Guna Seetharaman†, Kannappan Palaniappan*
* Department of Computer Science, University of Missouri-Columbia, MO 65211, USA
† Advanced Computing Concepts, U.S. Naval Research Laboratory, Washington, DC 20375, USA
‡ Department of Computer Science, University of Technology, Baghdad, Iraq
Email: * nmahyd@mail.missouri.edu, *{bunyak, palaniappank}@missouri.edu, †Gunasekaran.Seetharaman@rl.af.mil

Abstract—Recently our group proposed LoFT (Likelihood of Features Tracking) tracker system [1] that can successfully track objects of interest under different scenarios of wide-area motion imagery and full motion video. LoFT is a recognition-based single target tracker that relies on fusion of multiple complementary features. In this paper, LoFT is extended with a kernelized correlation filter (KCF) module to incorporate a robust continuous target template update scheme to better localize the target and to recover from sudden appearance changes and occlusions. Decision module using peak-to-sidelobe ratio is added to KCF module to prevent error accumulation from blending non-target regions to target template during update, and to prevent fusion of the KCF likelihood map to the other LoFT feature likelihood maps when the KCF response is not reliable. KC-LoFT is a single object tracker that fuses the most discriminative features from LoFT and KCF to better localize the target object in the search window. KC-LoFT was tested on ABQ aerial wide area motion imagery dataset [2] and produced promising results compared to recent state-of-the-art tracking systems in term of accuracy and robustness.

Index Terms—Tracking, wide area motion imagery, kernelized correlation filter, discriminative modeling, ridge regression.

I. INTRODUCTION

Visual object tracking is one of the most essential problems in computer vision. Wide variety of applications including visual surveillance, video summarization, sports video analysis, and biomedical video analysis rely heavily on visual object tracking. Tracking is the process of estimating the locations of objects of interest over time. Despite all the recent advancements in computer vision, visual tracking remains to be a challenging task because of the real-world conditions such as partial or full occlusions, background clutter, shadow and other illumination artifacts. Many visual tracking approaches have been proposed in the literature to address the unique requirements of different applications. These tracking approaches can be categorized in various ways such as detection-based [3] [4] versus recognition-based [5] [6]; single-object [6] [7] [8] versus multi-object [9] [10] [11]; or according to how the objects are localized within the search region, as generative [12] [13] versus discriminative methods [14] [15].

Correlation filter-based tracking is a discriminative method in which a filter based on the object appearance is used to estimate the most likely target location in the search window by distinguishing the target from its surrounding background. Convolution of the search window image with the filter is performed in Fourier domain as element-wise multiplications to reduce complexity. The used filters are updated online to adapt to object appearance changes. Since their first use in visual tracking [15] [16], correlation filter-based tracking has been extended in multiple ways. [14] [17] improved adaptation to scale changes; while [18] incorporated histogram based ridge regression learning to improve robustness to fast deformation and shape variations.

For a tracking system to handle a variety of challenging conditions, it is important to use multiple information cues robust to these conditions. Feature fusion needs to be explored in order to overcome weakness of individual information cues and to strength the overall process. Some feature fusion methods used in visual tracking are fixed linear combination [19] [20], adative weight update [21] [22], variance ratio [21]. Bayesian probability distribution [23], Bhattacharyya distance [24], peak-to-sidelobe ratio [15].

Our group developed a Likelihood of Features Tracking (LoFT) system [1] that fuses multiple sources of information about target and its environment to perform robust single object tracking. In this paper, we extend and improve LoFT by incorporating: (i) a discriminative module using kernelized correlation filters to strength LoFT’s adaptation to environment
The goal of correlation filter is to minimize the square error and used in the kernelized correlation filter (KCF) scheme. The new decision module uses peak-to-sidelobe ratio (PTSR) criterion to avoid fusing any irrelevant response from the kernelized correlation filters to the other LoFT feature maps and to prevent update of the regression parameters and correlation-based template during occlusion or other cases of sudden appearance change. We have tested the proposed KC-LoFT tracker on wide area motion imagery and compared its performance to non-deep learning trackers from VOT2015 [25] and VOT2016 [26] challenges (Table I) whose codes were publicly available.

The rest of this paper is organized as follows: Section II describes the original LoFT framework and its extension with kernelized correlation filters. Section III presents the experimental results. Section IV concludes the paper.

II. KC-LoFT: LIKELIHOOD OF FEATURES TRACKING WITH KERNELIZED CORRELATION

A. Likelihood of Features Tracking (LoFT) Framework

Likelihood of Features Tracking (LoFT) [1] is a recognition-based target tracking system initially designed for vehicle tracking in low frame rate wide area motion imagery, later also used in full motion video analysis. LoFT uses a rich feature set describing intensity, edge, shape and texture information. Target to search window feature comparison is performed using cross-correlation and sliding window histogram differencing using an efficient integral histogram computation scheme. The process produces a likelihood map for each individual feature. Different features are more sensitive or more robust against different target characteristics and environmental conditions. Fusing different feature likelihood maps enables adaptation of the tracker to scene dynamics and target appearance variability. LoFT uses two likelihood map weighting schemes, variance ratio (VR) [21] for histogram-based features and distractor index (DI) [27] for correlation-based features. LoFT appearance adaptation scheme maintains and updates a single template by calculating affine changes in the target to handle orientation and scale changes [1]. LoFT target template update is performed continuously to ensure accurate target localization. However continuous template update has two main problems: (1) drift when target template is updated with non-target features during partial or full occlusions; and (2) delayed recovery after occlusion events (many update steps are needed to realign target appearance).

B. KC-LoFT Kernelized Correlation Filter Module

In this paper we extend the original LoFT framework with kernelized correlation filter (KCF) [16] module. KC-LoFT kernelized correlation module uses two features, HoG and intensity, to localize the target within the search window. The two features are stacked to form a single vector \( x \) that is then used in the kernelized correlation filter (KCF) scheme. The goal of correlation filter is to minimize the square error over the sample \( x \) and the expected target \( y \) using the ridge regression loss function defined as follow

\[
\min_w \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|w\| \tag{1}
\]

where \( n \) is the length of the feature vector and \( \lambda \) controls the regularization parameter \( w \) over the linear combination of samples \( f(x) = w^T x \). Following the description and derivatives on [16] regression learning parameter, \( \alpha \) can be learned using

\[
\hat{\alpha} = \frac{\bar{y}}{\hat{k}^{xx} + \lambda} \tag{2}
\]

where \( \hat{k}^{xx} \) is a correlation kernel. Gaussian kernel is used as follow

\[
k^{xx'} = \exp\left(-\frac{1}{\sigma^2}(\|x\|^2 + \|x'\|^2) - 2F^{-1}(\hat{x} \odot \hat{x}^*) \right) \tag{3}
\]

Where \( \hat{x} \) is the DFT of \( x \), \( \hat{x}^* \) is the complex-conjugate of \( \hat{x} \), \( \odot \) is an element wise multiplication, and \( F \) is the discrete Fourier transform. To detect the position of the object, a new patch \( z \) (search window) is cropped form the location estimated from Kalman filter. The response is found according to the correlation between previously learned template \( \hat{x} \) and new patch \( z \).

\[
\hat{f}(z) = (\hat{k}^{xz}^*) \odot \hat{\alpha} \tag{4}
\]

The steps of the target detection and online training process are described in Algorithm 1, where \( \theta \) is the learning rate assumed to be 0.1.

C. KC-LoFT: Integration of Modules

KC-LoFT integrates the discriminative KCF module to the LoFT framework to enable robust and efficient localization of targets. KC-LoFT also incorporates an online template update
Algorithm 1 Target detection and training process using correlation filter module

**Input:** $P_{f-1}, z, \alpha_{f-1}, \tilde{x}$: predicted position, current patch (search region), dual space coefficient, the previous updated correlation-based template

**Output:** $P_f, \alpha_f, \tilde{x}_{new}$: estimated new position, the updated dual space coefficient, updated correlated-based template

1: Calculate the kernel $k_{\tilde{x}z} = \exp(-\frac{1}{2\sigma^2} (\|\tilde{x}\|^2 + \|z\|^2) - 2F^{-1}(\tilde{x} \odot \tilde{z}^*)$)
2: Calculate the response $\hat{f}(z) = (k_{\tilde{x}z})^* \odot \hat{\alpha}_{f-1}$
3: Find the new position $P_f$ from the maximum response $\hat{f}(z)$
4: Crop new patch $x_{new}$ with $P_f$ as a center of the region
5: Update the template $x_{new} = \tilde{x}z + (1 - \theta)x_{new}$
6: Calculate $\alpha_{new} = \frac{\theta}{k_{\tilde{x}z} + \lambda}$
7: Update dual space coefficient $\alpha_f = \theta \alpha_{f-1} + (1 - \theta) \alpha_{new}$

scheme to LoFT to prevent drifts and to enable robust handling of partial or full occlusions. Figure 2 illustrates modules involved in the proposed KC-LoFT pipeline.

Correlation filter learns filter parameters and updates template for each frame by including positive and negative examples within the search window using ridge regression. KC-LoFT fuses the responses from LoFT features with the correlation response from the correlation module to generate a final fused probability map that is used to localize the target. Because the correlation template update happens during the processing of every frame, online update on correlation-based template has faster response to appearance changes than appearance-based template. The response $\hat{f}(z)$ from the correlation filter module is fused with the other appearance-based template likelihood maps from LoFT to generate a final likelihood map to localize the expected position of the target. Peak-to-sidelobe ratio (PTSR) likelihood map evaluation criterion is used to avoid updating the correlation-based template and regression parameters during occlusion and to prevent correlation likelihood map to be fused with the remaining maps when it is unreliable. Peak-to-sidelobe ratio [15] evaluates the discrimination power of a likelihood peak as follow:

$$\text{PTSR} = \frac{P_{max} - \mu_{sidelobe}}{\sigma_{sidelobe}}$$

where $P_{max}$ is the value of the maximum response, $\mu_{sidelobe}$ and $\sigma_{sidelobe}$ are the mean and standard deviation of the likelihood map values except an $11 \times 11$ area around the maximum peak. Figure 4 shows sample likelihood maps and associated PTSR measurements.

During occlusion events, PTSR helps suspend correlation template update until the object appears again. Figure 3 illustrates KC-LoFT template update and likelihood fusion processes. Suspension of the template update process is important to preserve the target template through occlusion events.

### III. Experimental Results

Benchmark datasets and challenges are important for fair evaluation and comparison of trackers. Since 2013, VOT challenge group [28] has been organizing single object tracking challenges for selected full motion video datasets. Table I shows VOT2015 [25] and VOT2016 [26] ranks of LoFT and other state-of-the-art trackers used in this paper. All the listed trackers perform better than LoFT on these full motion video datasets. However, the nature of benchmark datasets and challenges is important for fair evaluation and comparison of trackers.
datasets can be one of the most important factors in tracker evaluation. LoFT and KC-LoFT are trackers developed for aerial wide area motion imagery. Aerial wide area motion imaginary (WAMI) datasets have different characteristics and challenges compared to the regular full motion videos. These WAMI videos suffers from extreme camera motion, low frame rate, frequent object deformations, rapid scale and appearance changes, small object sizes, shadow and illumination artifacts, partial and full occlusions.

In this paper, we have tested and evaluated the proposed KC-LoFT tracker and the selected state-of-the-art trackers on the ABQ aerial wide area motion imagery dataset for downtown Albuquerque, NM. We have manually generated ground-truth tracks for 136 cars within a 200 frames subset of the ABQ dataset. A frame from the ABQ dataset and zoomed views of three sample cars from the scene are shown in figure 5. For tracker performance evaluation, we have used two VOT challenge measures, **Accuracy** and **Robustness**. **Accuracy** measures how well the bounding box predicted by a tracker overlaps with the ground truth bounding box. **Robustness** measures number of times a tracker looses the target during tracking. Table II summarizes the tracking performances for the proposed KC-LoFT tracker and other state-of-the-art trackers on ABQ dataset. KC-LoFT increases both accuracy and robustness of the original LoFT tracker (by 9.6% and 5.1% respectively) and produces better or comparable results compared to the state-of-the-art trackers from the VOT2015 [25] and VOT2016 [26] challenges. Figure 6 shows sample tracking results for two cars. Trajectory color represents number of reinitializations after tracker failures. Lower number of trajectory colors indicates tracker robustness. In both cases KC-LoFT tracks the selected cars without any failures or restarts, while for the first car LoFT, DSST, and Staple require one or more restarts, and for the second car all the trackers except KC-LoFT require one or more restarts.

### IV. Conclusions

KC-LoFT is a recognition based object tracker that extends the Likelihood of Features Tracker (LoFT) framework with the kernelized correlation filter (KCF) scheme to better localize the target in a search window. KC-LoFT combines the LoFT strengths such as multiple complementary features robust to environmental conditions and shape deformations, with the KCF strength such as online template and parameter update and robustness to target shifts and rotations. KC-LoFT overcomes the LoFT and KCF limitations by improving the performance of correlation filter with object location and search window prediction, and by suspending update during occlusion events to keep the correlation-based template and parameters reliable. The use of appearance-based template helps KC-LoFT to better localize object during deformation, and online update helps adaptation to environmental changes.
Experimental results on wide area motion imagery show improved performance in term of accuracy and robustness compared to LoFT and better or comparable results compared to other state-of-the-art trackers.

REFERENCES


