Feature Fusion Using Ranking for Object Tracking in Aerial Imagery

Sema Candemir[a], Kannappan Palaniappan[a], Filiz Bunyak[a], Guna Seetharaman[b]

[a]University of Missouri, Dept. of Computer Science,Columbia, MO 65211, USA
[b]Air Force Research Laboratory, Rome, NY 13441, USA

ABSTRACT

Aerial wide-area monitoring and tracking using multi-camera arrays poses unique challenges compared to standard full motion video analysis due to low frame rate sampling, accurate registration due to platform motion, low resolution targets, limited image contrast, static and dynamic parallax occlusions.[1–3] We have developed a low frame rate tracking system that fuses a rich set of intensity, texture and shape features, which enables adaptation of the tracker to dynamic environment changes and target appearance variabilities. However, improper fusion and overweighting of low quality features can adversely affect target localization and reduce tracking performance. Moreover, the large computational cost associated with extracting a large number of image-based feature sets will influence tradeoffs for real-time and on-board tracking. This paper presents a framework for dynamic online ranking-based feature evaluation and fusion in aerial wide-area tracking. We describe a set of efficient descriptors suitable for small sized targets in aerial video based on intensity, texture, and shape feature representations or views. Feature ranking is then used as a selection procedure where target-background discrimination power for each (raw) feature view is scored using a two-class variance ratio approach. A subset of the k-best discriminative features are selected for further processing and fusion. The target match probability or likelihood maps for each of the k features are estimated by comparing target descriptors within a search region using a sliding window approach. The resulting k likelihood maps are fused for target localization using the normalized variance ratio weights. We quantitatively measure the performance of the proposed system using ground-truth tracks within the framework of our tracking evaluation test-bed that incorporates various performance metrics. The proposed feature ranking and fusion approach increases localization accuracy by reducing multimodal effects due to low quality features or background clutter. Adaptive feature ranking increases the robustness of the tracker in dynamically changing environments especially when the object appearance is changing.

1. INTRODUCTION

Visual tracking in wide-area motion imagery (WAMI) is an important component in processing, exploitation and dissemination (PED) of intelligence, surveillance and reconnaissance (ISR) information. However, tracking in wide-area aerial imagery poses a number of difficulties[1–8] for traditional vision-based tracking algorithms due to low frame rate sampling, large image size, low resolution targets, limited target contrast, foreground clutter (similar targets), background clutter, severe shadow or geometric occlusion areas, static and dynamic parallax occlusions, camera motion, registration and mosacing across multiple cameras, etc. These difficulties make the tracking task in aerial imagery more challenging compared to standard ground-based or even narrow field-of-view full motion video analysis. Figure 1 shows sample single camera images from the CLIF 2007 WAMI dataset.

Figure 1. Sample Columbus Large Image Format (CLIF 2007)[9] wide-area contrast enhanced images over OSU.

Palaniappan *et al.*[2,10] developed a target tracking system suitable for low frame rate aerial motion imagery. The system models the target using a set of intensity, texture and shape feature descriptors. It computes the target match likelihoods for each feature descriptor by comparing the target to a local search image region or area of interest using a sliding window approach. The likelihood images for all descriptors are fused using a weighted sum. Local maxima in the fused likelihood map that exceeds a predetermined threshold are considered as potential target locations. Figure 2 summarizes the general structure of the tracking system relevant to feature ranking, selection and fusion; the filtering and prediction feedback as well as appearance update modules[2] for example are not shown.

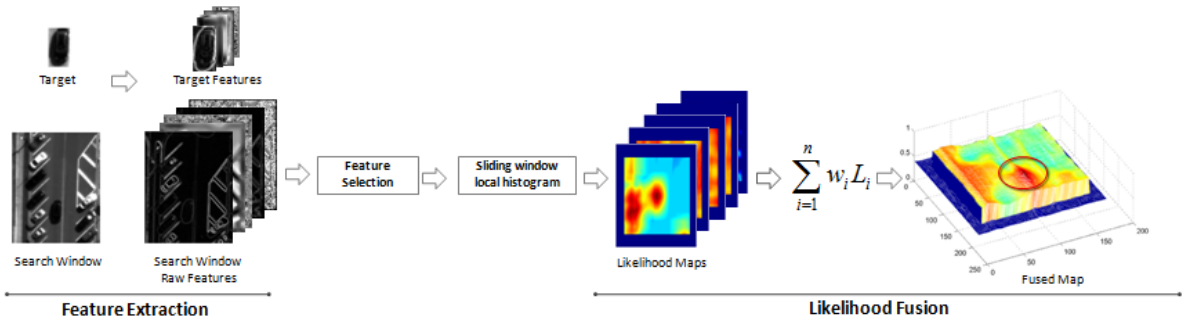


Figure 2. Overview of feature extraction and fusion process. Feature probability maps are fused using a weighted sum and used for target localization. L_i represents the likelihood of feature i, w_i represents the weight of feature map L_i.

Using multiple features provides more robust localization especially in cluttered environments. It also enables adaptation of the tracker to dynamic environment changes and target appearance variabilities. However, each feature performs differently depending on the target characteristic and environment changes. So equally weighted fusion of likelihood maps may decrease performance, if some of the features under perform in that environment. Additionally combining all features, even with suitable weights, may cause multi-modal effects in the fused likelihood maps. Another benefit of feature selection is computational cost savings and storage requirements to facilitate real-time and onboard tracking applications. Figure 3 shows an example of improved car localization performance using the top three ranked features compared to equal weight fusion.

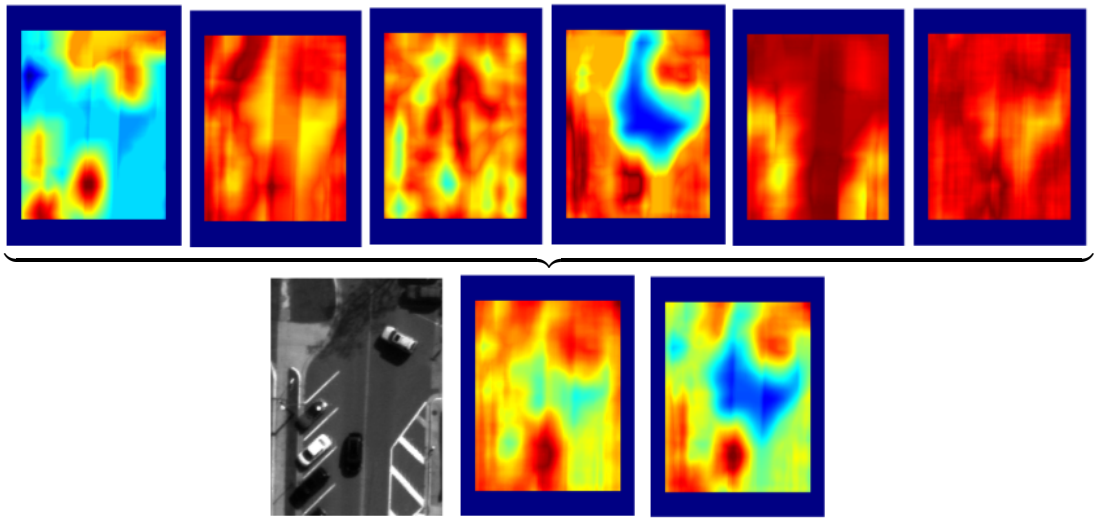


Figure 3. **Top Row:** Target match likelihood maps obtained using intensity histograms, gradient magnitude histogram, shape index, normalized curvature index, histogram of oriented gradients (HoG) descriptor, local binary pattern (LBP) histogram. **Bottom Row:** Region of interest search window (left), equal weight fusion with probability on target of 0.895 (middle), ranked feature fusion (best 3 features) with match probability on target of 0.946 (right).

In this paper we explore the two-class variance ratio-based[11] feature ranking and feature fusion in the context of tracking in wide-area motion imagery. We apply a dynamic feature ranking algorithm to our tracker system in order to improve the target localization in the fused likelihood map and increase the tracking performance. We incorporate the variance ratio calculation into our tracking system as a feature selection and linear weighted sum fusion method. Raw tuned features are ranked according to their separability power such that if the feature has high separability then it's variance ratio is expected to be higher. The k-best features (higher variance ratio) are selected for fusion. The tracker system fuses the likelihood maps of selected features through a weighted sum. The feature ranking and fusion scheme increases the localization accuracy by reducing multi-modal effects introduced by low quality features and background clutter. We quantitatively measure the performance of the described system with respect to ground-truth detections and tracks using our tracking evaluation test-bed that incorporates various experiments and performance metrics.

Section 2 explains the variance ratio feature ranking and fusion method. Section 2.1 discuss the pros and cons of the variance ratio method including variance ratio behavior with heterogenous feature descriptors and varying peak strength and shape. Section 3 gives experimental results of the tracking system followed by conclusions.

2. VARIANCE RATIO FEATURE RANKING, SELECTION AND WEIGHTING

Feature selection methods[12,13] choose relevant features among the feature candidates in order to improve the classification or detection performance. In the context of tracking, online feature selection is a powerful approach to dynamically maximize separability of target from background regions and improve target localization. In this work, we choose an online feature selection mechanism which selects the most relevant feature set for each frame during tracking. Collins *et al.*[11,14] state that the best features for tracking an object are the features that are the most discriminative between target and background at a particular time. They adaptively weight the features according to its discriminative power between target and background using the two-class variance ratio measure. We adopted this idea and applied it to our tracking system using tuned raw features.[11] We measure the target and background discrimination power for each feature. Target and background pixels are sampled from the location of the tracked object at previous frame. Figure 4 indicates the target (red) and background (blue) regions on several features. The target region is the bounding box which contains the tracked object and the surrounding square annulus contains background pixels. The graphs below the images show the feature distributions of foreground and background regions. In order to calculate the variance ratio, we first measures the log likelihood ratio L between foreground and background regions for each (tuned) feature i using Eq. 1,

$$L(i) = log\frac{max(fg(i), \delta)}{max(bg(i), \delta)} \tag{1}$$

where $fg(i)$ and $bg(i)$ are the discrete probability distributions of target and background regions, δ= 0.0001 is a small value that prevents division by zero. Then variance ratio of each feature i is calculated using Eq. 2.

$$VR(L_i, fg, bg) = \frac{var(L_i; (fg + bg)/2)}{var(L_i, fg) + var(L_i, bg)}. \tag{2}$$

The variance ratio between foreground and background histograms scores is used to measure the discriminative power of each feature. A higher ratio indicates higher discrimination between the target and background. Finally, features are ranked according to their foreground-background discriminative power and the k best features are used to compute the fused likelihood map.

There are many potential approaches for robust feature fusion.[15–18] Our system uses a weighted sum approach which is more robust than several other fusion methods.[19] In a weighted sum approach, the challenging parameter to estimate is the relative importance of the fused modules. In this work, we use the normalized variance ratio scores associated with the selected features as the relative weights for fusion (Eq.3).

$$w_i \approx \frac{VR(L_i; fg, bg)}{\sum_i^n VR(L_i; fg, bg)} \tag{3}$$

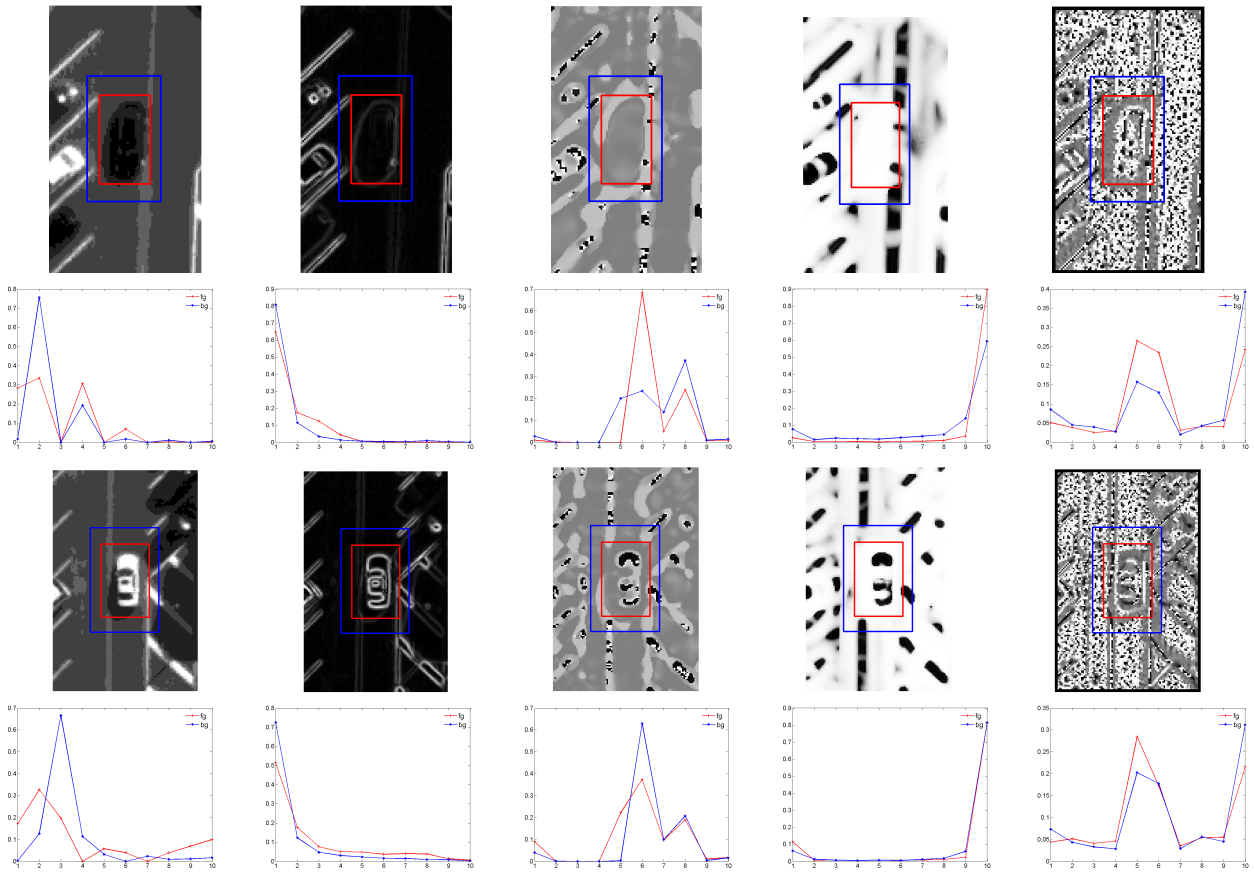


Figure 4. Raw features and regional histograms for foreground and background regions of white and black car. Features: Intensity, gradient magnitude, Hessian shape index, normalized curvature index, and local binary pattern features

Online feature selection provides two main advantages for our tracking system: (i) We adaptively select the most relevant features for every frame instead of relying on all features and dynamically update this set over time so that irrelevant features are automatically discarded during tracking in a context sensitive way. (ii) Instead of using equal weights for fusion, we adaptively compute weights for each feature, which increases the robustness of the tracker in dynamically changing environments.

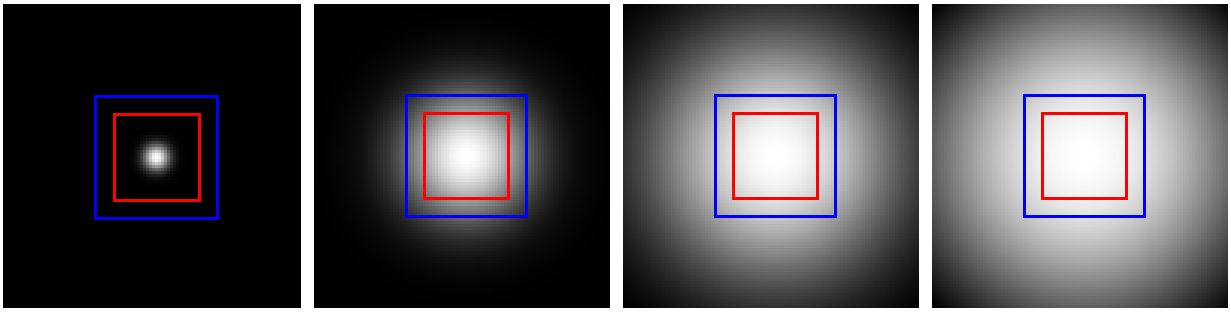
2.1 Behavior of Variance Ratio and Peak Shape for Heterogeneous Descriptors

The variance ratio has some limitations when applied to features with narrow peak shapes which we discuss in more detail in this section. Additional features that incorporate local regional or target information can be used to improve the feature detection performance.[2,10,14] Some of these likelihood maps can exhibit narrower peaks especially correlation type features. Using the variance ratio to fuse features with different peak variance/spread behavior can lead to underweighting. For example, Figure 5 shows a narrow peaky feature with high contrast but a low value for the variance ratio as shown in Figure 6. The broader peaked distributions have a higher variance ratio and would consequently receive a higher weight which is undesirable.

The variance ratio based ranking approach adaptively selects the relevant features; it is more suitable when all of the features or descriptors have similar (peak shape) characteristics. However, a tracking system may incorporate heterogeneous types of features or descriptors (i.e. correlation, histogram or tuned).[10] We analyze the variance ratio behavior on synthetic feature distributions to demonstrate the limitation of the variance ratio. We simulate feature probability maps using a single Gaussian distribution with different σ values for the peak shape. These experiments indicate that features with a narrow sharp peak shape have a lower variance ratio

compared to features with a broad peak shape. This unfavorable behavior of the variance ratio is shown in the graph in Figure 6 which includes a plot of the foreground and background means and variances.

The variance ratio is suitable for the (raw) tuned feature maps since they usually do not exhibit highly narrow peaky distributions. If heterogeneous descriptors are used with narrow and broad peaks then the distractor index approach would be a better measure for feature ranking.[2,10]



$\sigma = 3$ $\sigma = 10$ $\sigma = 30$ $\sigma = 50$

Figure 5. Feature/descriptor simulation using single Gaussian distributions. The foreground region is shown in red and the background region is between the red and blue rectangular regions. The variation in peak shape is captured through the standard deviation of the Gaussian.

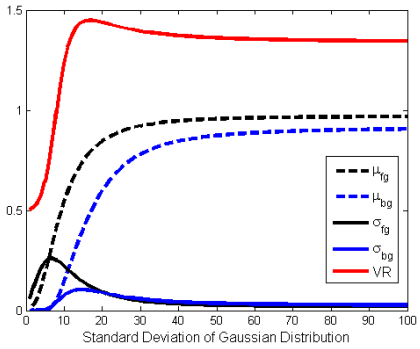


Figure 6. Plot of variance ratio versus standard deviation of the peak shape showing that narrow sharp peaks have smaller variance ratio than broad peaks leading to a lower ranking of a more discriminative feature. The dashed curves are the means and the solid curves are the variances for the foreground (black color) and background (blue color) regions.

3. EXPERIMENTS

We evaluated the performance of dynamic feature evaluation and fusion method on CLIF[9] dataset under challenging background conditions such as occlusions from shadows, turning vehicles, low contrast and fast vehicle motion. Figure 7 shows sample registered CLIF images from the published test dataset for 14 cars across 12 sequences.[7,9] We manually adjusted the target (unregistered) ground-truth in the post-registered sequences.

In order to show that feature selection and adaptive weight fusion improves the target localization, we setup a tracking test bed system (Figure 2). The system first calculates the variance ratio of features using the ground truth locations. Then, it ranks features according to their variance ratio scores, and selects the k best feature. In the next frame, the selected features are fused using the variance ratio weights from the previous frame. We calculate the target probabilities and peak localization using the fused likelihood map. All local maxima which exceed an experimentally determined threshold is considered to be a potential target location. For the quantitative evaluation of detection performance, we measure the target probability of equal weight fusion,[19] VR-weight fusion,[11] and Rank-k fusion methods. Except for the fusion scheme, all other testing parameters are kept the same. Higher probability on the target is indicative of better fusion. Figures 8 and 9 show the temporal

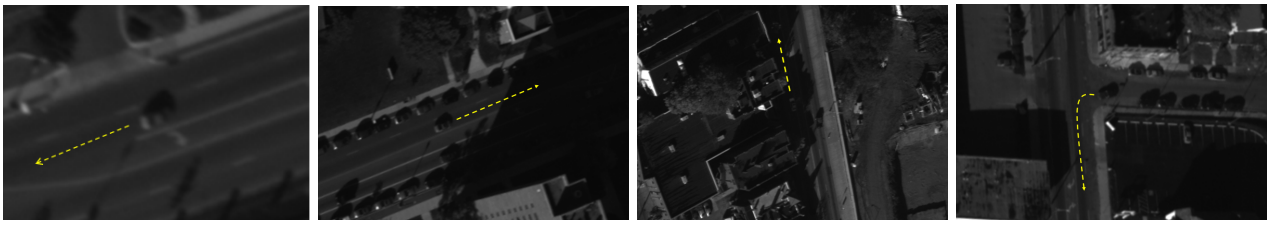


Figure 7. Sample original images with vehicle tracks showing challenging conditions: Low foreground-background contrast (C0-3-0), Large shadow region (C0-3-0), Geometric and shadow occlusion (C2-4-1), Turning target (C4-1-0).

variation in target location probabilities for four CLIF sequences using eight different fusion methods – EqualW fuses all feature likelihood maps using equal weights, VR-All fuses all likelihoods using VR weights, VR-Rank-k fuses just the k best feature likelihood maps. In the four cases shown, VR-Rank-1 and VR-Rank-2 methods produce higher peaks on the target with a better ranking which improves the performance of target localization. In Figure 8, the regions between frames 8-13 and frames 38-41 are occluded areas where the search regions are almost black so individual likelihood maps have low probabilities.

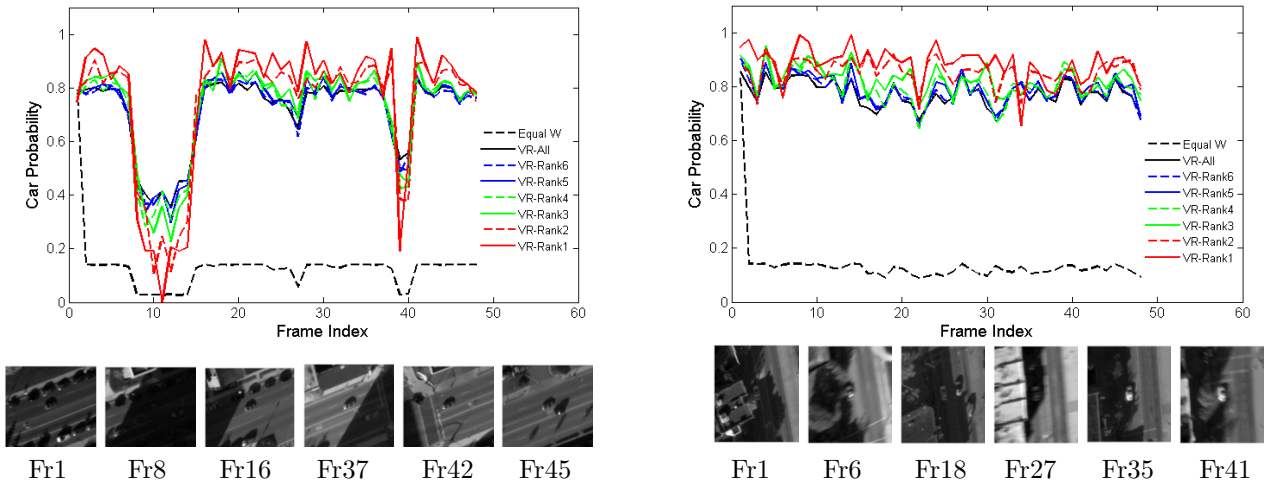


Figure 8. **Top:** Target probability graphs of feature fusion methods for C0-3-0 (Left) and C2-4-1 (Right) with tracks shown in Fig. 7. **Bottom:** Sample search regions in corresponding frames.

Another evaluation method is to rank the local peaks in the fused likelihood maps to assess performance – the lower the rank of the target peak the better the performance. The maximum peak is assigned the highest rank (number 1). If the highest rank is on the target, it indicates better fusion. Figure 10 shows the ranked peaks for Equal weights, VR-All weights, VR-Rank-5 and VR-Rank-3 for a sample region of interest. Note that the highest peak in Equal Weight and VR-All methods are located on another vehicle. On the other hand, the highest peak using VR-Rank-5 and VR-Rank-3 are correctly associated with the target.

Figure 11 shows performance based on average recall versus number of selected peaks for each fusion method over all frames in the test sequences. Recall is the ratio of the correctly tracked frames to the ground truth and is defined as,

$$\text{Recall} = \frac{\#\text{ Correct Frames}}{\#\text{ Ground Truth Frames}} \tag{4}$$

The 12 test sequences includes 14 vehicles with complex appearance and background environments.[7] Each vehicle has approximately 50 frames in the video. It can be seen that ranking features for fusion (VR-Rank-1 and VR-Rank-2) tends to produce better recall compared to the Equal Weight and VR-All methods. The significant peak recall performance difference compared to the results reported previously using PSS imagery[2] is due to the selection of lower contrast vehicles in the CLIF ground-truth and the presence of more shadows and occlusions along the target tracks in the CLIF sequences. The peak-recall performance with CLIF data could be

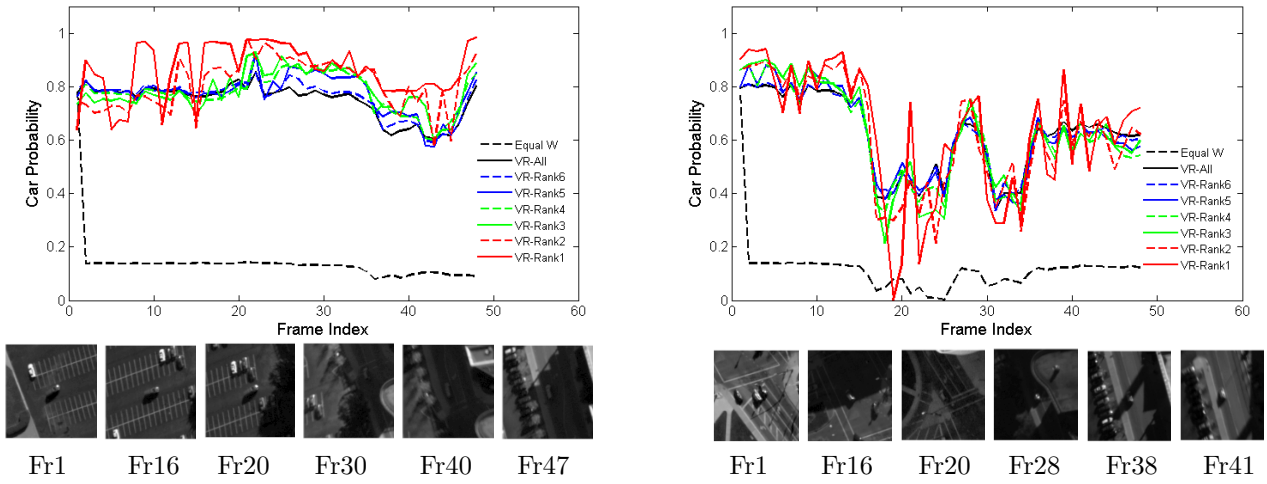


Figure 9. **Top:** Target probability graphs of feature fusion methods for C1-4-6 (Left) and C1-4-0 (Right). **Bottom:** Sample search regions in corresponding frames.

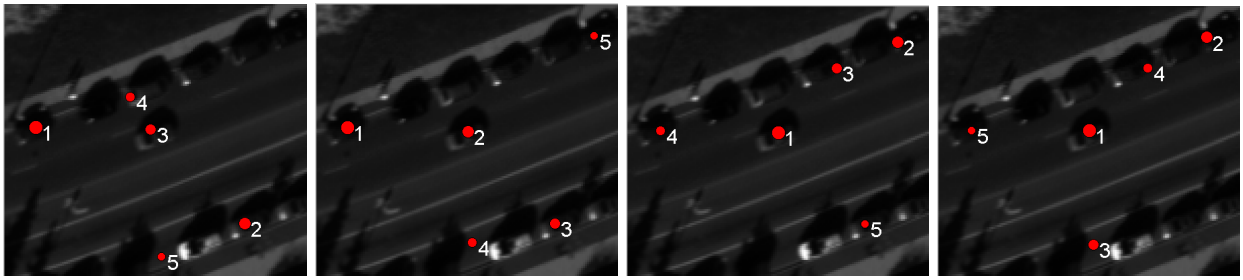


Figure 10. Ranked peaks of fused maps of EqualWeight, VR-All, VR-Rank-5, and VR-Rank-3 for the same region.

improved using a number of different approaches including additional spatial feature descriptors such as median binary patterns[20] or adaptive robust structure tensors,[21] motion descriptors,[22] improved registration,[8,23] spatial relationships,[24] and localized image segmentation methods.[16,17,25–27]

4. CONCLUSIONS AND FUTURE WORK

This paper presents a framework for dynamic feature evaluation and ranking-based fusion for vehicle tracking in aerial imagery. Ranking is based on ordering (raw) features according to their foreground and background separability in an adaptive context sensitive manner for each frame during the course of tracking where the object appearance can be changing. Foreground and background separability was measured using a two-class variance ratio. Other ranking measures such as the distractor index can be used and the behavior of the variance ratio needs further study since the variance ratio for features with narrow sharp peaks (in the a posteriori probabilty estimates) can result in an underestimate of the feature rank or weight. Experiments show that online ranking-based feature selection improves target localization and overall object tracking performance. Feature fusion using dynamic feature ranking-based selection consistently outperforms equal weighting based feature fusion (using all features). Using a subset of features selected through ranking is not only superior to feature fusion using the full feature set but also improves computational efficiency for real-time applications and can be incorporated in parallel video processing implementations.[28–30]

ACKNOWLEDGMENTS

This research was partially supported by U.S. Air Force Research Laboratory (AFRL) under agreement AFRL FA8750-11-C-0091. Approved for public release (case 88ABW-2012-2535). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies,

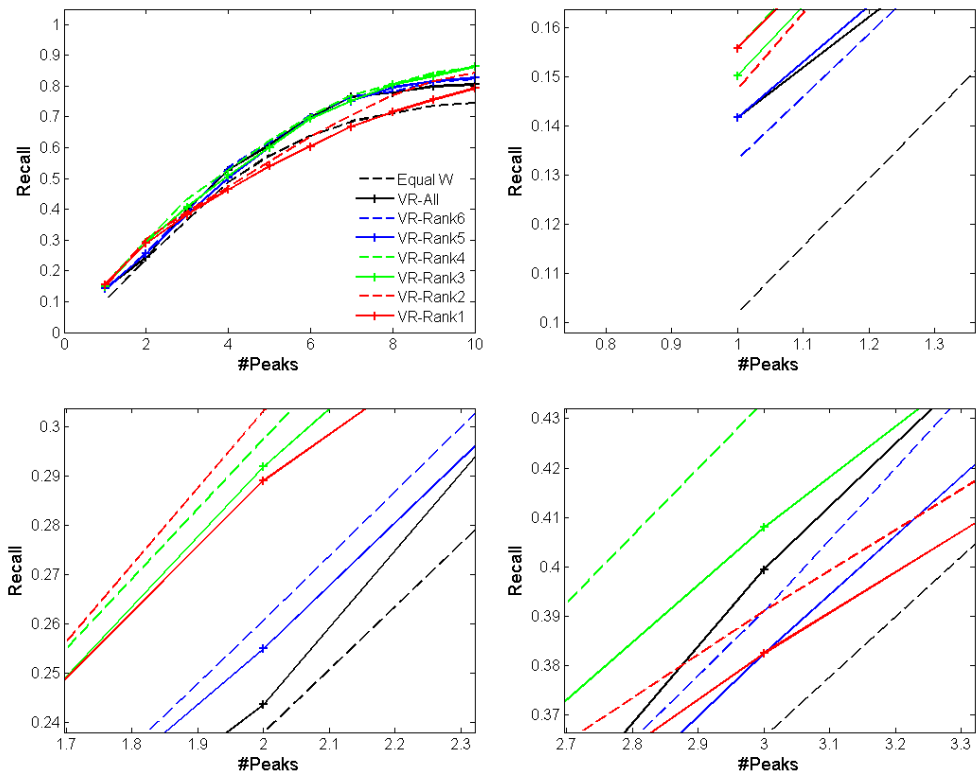


Figure 11. Average recall versus number of selected peaks for each fusion method over all frames for 14 vehicles (showing zoomed up parts of the full graph shown in top left).

REFERENCES

[1] Palaniappan, K., Rao, R., and Seetharaman, G., "Wide-area persistent airborne video: Architecture and challenges," in [*Distributed Video Sensor Networks: Research Challenges and Future Directions*], Banhu, B., Ravishankar, C. V., Roy-Chowdhury, A. K., Aghajan, H., and Terzopoulos, D., eds., ch. 24, 349–371, Springer (2011).

[2] Palaniappan, K., Bunyak, F., Kumar, P., Ersoy, I., Jaeger, S., Ganguli, K., Haridas, A., Fraser, J., Rao, R., and Seetharaman, G., "Efficient feature extraction and likelihood fusion for vehicle tracking in low frame rate airborne video," in [*13th Int. Conf. Information Fusion*], (2010).

[3] Haridas, A., Pelapur, R., Fraser, J., Bunyak, F., and Palaniappan, K., "Visualization of automated and manual trajectories in wide-area motion imagery," in [*15th Int. Conf. Information Visualization*], 288–293 (2011).

[4] Fraser, J., Haridas, A., Seetharaman, G., Rao, R., and Palaniappan, K., "KOLAM: An extensible cross-platform architecture for visualization and tracking in wide-area motion imagery," in [*Proc. SPIE Conf. Geospatial InfoFusion II (Defense, Security and Sensing: Sensor Data and Information Exploitation)*], **8396** (2012).

[5] Blasch, E., Deignan, P., Dockstader, S., Pellechia, M., Palaniappan, K., and Seetharaman, G., "Contemporary concerns in geographical/geospatial information systems (GIS) processing," in [*Proc. IEEE National Aerospace and Electronics Conference (NAECON)*], 183–190 (2011).

[6] Porter, R., Fraser, A. M., and Hush, D., "Wide-area motion imagery," *IEEE Signal Processing Magazine* **27**(5), 56–65 (2010).

[7] Ling, H. and *et al.*, "Evaluation of visual tracking in extremely low frame rate wide area motion imagery," in [*14th Int. Conf. on Information Fusion*], 1866–1873 (2011).
[8] Hafiane, A., Palaniappan, K., and Seetharaman, G., "UAV-video registration using block-based features," in [*IEEE Int. Geoscience and Remote Sensing Symposium*], **II**, 1104–1107 (2008).
[9] AFRL CLIF dataset OSU, "https://www.sdms.afrl.af.mil/index.php?collection=clif2007," (2007).
[10] Pelapur, R., Candemir, S., Poostchi, M., Bunyak, F., Wang, R., Seetharaman, G., and Palaniappan, K., "Persistent target tracking using likelihood fusion in wide-area and full motion video sequences," in [*15th Int. Conf. Information Fusion*], (2012).
[11] Collins, R., Liu, Y., and Leordeanu, M., "Online selection of discriminative tracking features," *IEEE Trans. PAMI* **27**, 1631–1643 (2005).
[12] Maggio, E. and Cavallaro, A., [*Video Tracking: Theory and Practice*], Wiley (2011).
[13] Poostchi, M., Bunyak, F., and Palaniappan, K., "Feature selection for appearance-based vehicle tracking in geospatial video," in [*Proc. SPIE Conf. Geospatial InfoFusion II (Defense, Security and Sensing: Sensor Data and Information Exploitation)*], **8396** (2012).
[14] Yin, Z., Porikli, F., and Collins, R., "Likelihood map fusion for visual object tracking," in [*IEEE Workshop Appl. Comput. Vis.*], 1–7 (2008).
[15] Julier, S. J., Uhlmann, J. K., Walters, J., Mittu, R., and Palaniappan, K., "The challenge of scalable and distributed fusion of disparate sources of information," in [*SPIE Proc. Multisensor, Multisource Information Fusion: Architectures, Algorithms and Applications*], **6242**(1), Online (2006).
[16] Bunyak, F. and *et al.*, "Geodesic active contour based fusion of visible and infrared video for persistent object tracking," in [*IEEE Workshop App. of Computer Vision*], (2007).
[17] Bunyak, F., Palaniappan, K., Nath, S. K., and Seetharaman, G., "Flux tensor constrained geodesic active contours with sensor fusion for persistent object tracking," *J. Multimedia* **2**, 20–33 (August 2007).
[18] Maggio, E., Smeraldi, F., and Cavallaro, A., "Adaptive multi-feature tracking in a particle filtering framework," *IEEE Trans. on Circuits and Systems for Video Technology* **17**, 1348–1359 (2007).
[19] Kittler, J., Hatef, M., Duin, R., and J, M., "On combining classifiers," *IEEE Trans. PAMI* **20(3)**, 226–239 (March 1998).
[20] Hafiane, A., Seetharaman, G., Palaniappan, K., and Zavidovique, B., "Rotationally invariant hashing of median patterns for texture classification," *Lecture Notes in Computer Science (ICIAR)* **5112**, 619–629 (2008).
[21] Nath, S. and Palaniappan, K., "Adaptive robust structure tensors for orientation estimation and image segmentation," *Lecture Notes in Computer Science (ISVC)* **3804**, 445–453 (2005).
[22] Palaniappan, K., Ersoy, I., and Nath, S. K., "Moving object segmentation using the flux tensor for biological video microscopy," *Lecture Notes in Computer Science (PCM)* **4810**, 483–493 (2007).
[23] Seetharaman, G., Gasperas, G., and Palaniappan, K., "A piecewise affine model for image registration in 3-D motion analysis," in [*IEEE Int. Conf. Image Processing*], 561–564 (2000).
[24] Nath, S. K., Palaniappan, K., and Bunyak, F., "Accurate spatial neighborhood relationships for arbitrarily-shaped objects using Hamilton-Jacobi GVD," *Lecture Notes in Computer Science (SCIA)* **4522**, 421–431 (2007).
[25] Nath, S. K., Palaniappan, K., and Bunyak, F., "Cell segmentation using coupled level sets and graph-vertex coloring," *Lecture Notes in Computer Science (MICCAI)* **4190**, 101–108 (2006).
[26] Bunyak, F. and Palaniappan, K., "Efficient segmentation using feature-based graph partitioning active contours," in [*12th IEEE Int. Conf. Computer Vision*], 873–880 (2009).
[27] Nath, S. K. and Palaniappan, K., "Fast graph partitioning active contours for image segmentation using histograms," *EURASIP Journal on Image and Video Processing* , 9p (2009).
[28] Kumar, P., Palaniappan, K., Mittal, A., and Seetharaman, G., "Parallel blob extraction using the multi-core Cell processor," *Lecture Notes in Computer Science (ACIVS)* **5807**, 320–332 (2009).
[29] Mehta, S., Misra, A., Singhal, A., Kumar, P., Mittal, A., and Palaniappan, K., "Parallel implementation of video surveillance algorithms on GPU architectures using CUDA," in [*17th IEEE Int. Conf. Advanced Computing and Communications (ADCOM)*], (2009).
[30] Bellens, P., Palaniappan, K., Badia, R. M., Seetharaman, G., and Labarta, J., "Parallel implementation of the integral histogram," *Lecture Notes in Computer Science (ACIVS)* **6915**, 586–598 (2011).