Robust Object Matching for Persistent Tracking with Heterogeneous Features

Yanlin Guo, *Member*, *IEEE*, Steve Hsu, *Member*, *IEEE*, Harpreet S. Sawhney, *Member*, *IEEE*, Rakesh Kumar, *Member*, *IEEE*, and Ying Shan, *Senior Member*, *IEEE*

Abstract—This paper addresses the problem of matching vehicles across multiple sightings under variations in illumination and camera poses. Since multiple observations of a vehicle are separated in large temporal and/or spatial gaps, thus prohibiting the use of standard frame-to-frame data association, we employ features extracted over a sequence during one time interval as a vehicle fingerprint that is used to compute the likelihood that two or more sequence observations are from the same or different vehicles. Furthermore, since our domain is aerial video tracking, in order to deal with poor image quality and large resolution and quality variations, our approach employs robust alignment and match measures for different stages of vehicle matching. Most notably, we employ a heterogeneous collection of features such as lines, points, and regions in an integrated matching framework. Heterogeneous features are shown to be important. Line and point features provide accurate localization and are employed for robust alignment across disparate views. The challenges of change in pose, aspect, and appearances across two disparate observations are handled by combining a novel feature-based *quasi-rigid* alignment with *flexible* matching between two or more sequences. However, since lines and points are relatively sparse, they are not adequate to delineate the object and provide a comprehensive matching set that covers the complete object. Region features provide a high degree of coverage and are employed for continuous frames to provide a delineation of the vehicle region for subsequent generation of a match measure. Our approach reliably delineates objects by representing regions as robust blob features and matching multiple regions to multiple regions using Earth Mover's Distance (EMD). Extensive experimentation under a variety of real-world scenarios and over hundreds of thousands of Confirmatory Identification (CID) trails has demonstrated about 95 percent accuracy in vehicle reacquisition with both visible and Infrared (IR) imaging cameras.

Index Terms—Video object tracking and reacquisition, object matching, feature matching, image alignment and matching.

1 Introduction

Object tracking from aerial platforms requires data association over long periods of time. The object of interest, vehicles for the purposes of this paper, may not remain in the field of view continuously through the course of tracking. The tracked objects leave the field of view because of occlusions and inaccuracies in platform pointing directions. When the vehicles appear again, the tracker needs to verify if the currently observed vehicles are indeed the same as the ones being tracked earlier. Another important visual surveillance task requires multiple observations of the same vehicle viewed from different spatial sightings to be reliably associated. In both applications, we need to compute matching scores between a model (learning) sequence and a query sequence, assuming that frame-to-frame tracking is given as input. Several representative learning and query "object chips" are shown in Fig. 1. It is obvious that standard frame-to-frame association techniques cannot be directly applied to match the learning and query sequences in these applications because of the amount of object scale, pose and appearance change, the background clutter, and the lack of temporal and spatial continuity.

Despite a flurry of research on object matching and recognition [1], [2], [3], [4], [5], [6], [7], [8], online object fingerprinting still remains a very challenging problem because of the following reasons:

1. Limited training data is available. In contrast with traditional approaches to object identification in visual imagery, we cannot assume that every object has been modeled beforehand.
2. There can be drastic pose changes between the learning and query sequences. It is difficult to find reliable invariant feature representations because of occlusion and aspect change.
3. There can be large appearance changes. The presence of shadow and specularity makes matching even more challenging.
4. Video objects captured from various platforms and resolutions (2-20 cm/pixel typically) have to be handled.
5. It is not realistic to require that the object be accurately segmented from the background, thus object masks may not be accurate.
6. There may be multiple similar objects present at the same time.

To match objects under large pose, scale, and appearance changes and with background clutter and confusers, it is crucial to utilize as much information as possible. In this paper, we propose a novel object matching technique based on the exploitation and combination of heterogeneous

- *Y. Guo, H.S. Sawhney, R. Kumar, and Y. Shan are with the Sarnoff Corporation, 201 Washington Road, CN5300, Princeton, NJ 08543. E-mail: {yguo, hsawhney, rkumar, yshan}@sarnoff.com.*
- *S. Hsu is with Canesta, Inc., 965 West Maude Avenue, Sunnyvale, CA 94085. E-mail: shsu@canesta.com.*

Manuscript received 4 Oct. 2005; revised 3 Apr. 2006; accepted 31 July 2006; published online 18 Jan. 2007.
Recommended for acceptance by S. Baker.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0537-1005.
Digital Object Identifier no. 10.1109/TPAMI.2007.1052.

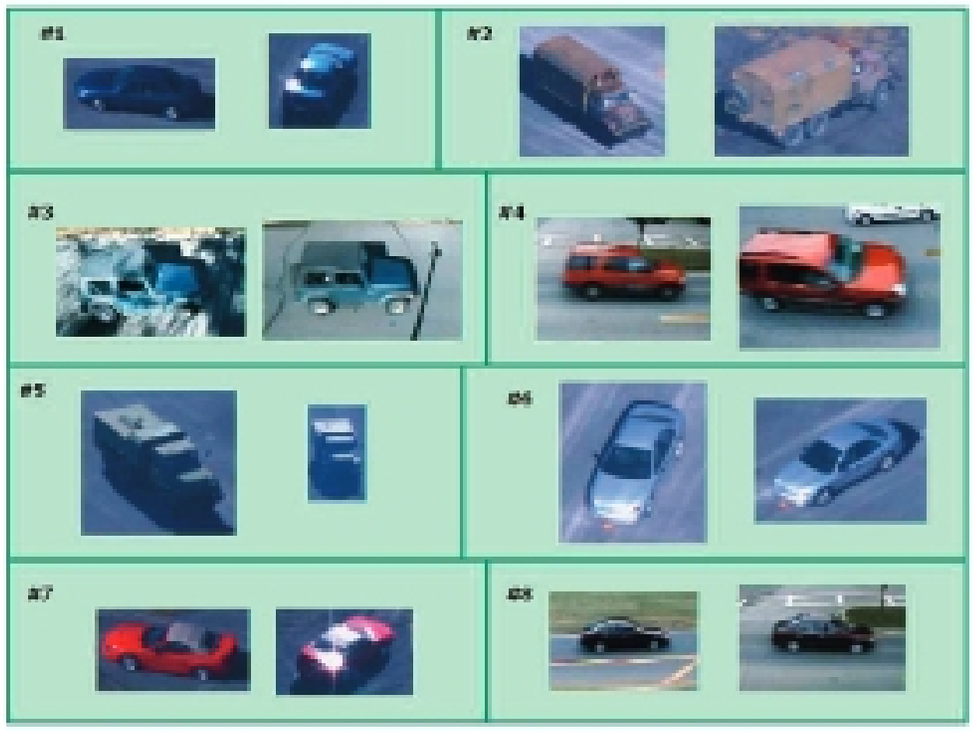

Fig. 1. Some representative object matching examples. Objects separated by a temporal or spatial gap from aerial and ground platforms are required to be matched against each other.

features: corner-like and line features for reliable geometric alignment and blob-like region features for comprehensive coverage, delineation, and matching of the object region. Each feature type is represented with suitable unique invariant representations and plays a different role in matching object geometry, appearance, and topology. Specifically, blob-like features [9], [10] provide good coverage for an object, but they are usually not suitable for image alignment due to the lack of localization accuracy. However, they can be consistently tracked across frames over a short period of time. Utilizing blob-like features *within* a sequence provides an accurate object mask for subsequence object matching across a sequence, if appropriate region descriptors and matching criteria are utilized [11], [12]. Outliers such as background clutter can be eliminated in the process of region matching. In addition, blob features can be used for overall object part configuration (topology) matching. Corner-like features and line features cannot provide sufficient extent of object coverage, but they possess good localization property and they are effective for object geometry matching (alignment), especially in *cross* sequence matching (between query and learning sequences).

Key contributions of our approach are:

1. Development of heterogeneous feature descriptors and respective match measures that utilize corner, line, and region features in different stages of object matching.
2. Development of a framework of using "within-sequence matching" using region features to obtain precise and sufficient object coverage plus "across-sequence matching" with point and lines features to achieve accurate image alignment.
3. Development of a robust blob detector and a match metric (earth mover's distance, EMD) to effectively match and track regions.
4. Development of a quasi-rigid alignment method based on invariant corner and line features to align images under large pose and appearance changes. The method avoids the explicit computation of a nonparametric 3D motion field by approximating it with a feature constrained quasi-rigid piecewise parametric motion model. It does not need explicit camera calibration, or dense reconstruction of 3D scenes. It can handle both parametric and nonparametric motion models, which are suitable for video data captured from various platforms and resolutions.
5. Development of a novel flexible template matching scheme with entropy-based adaptive scale determination in oriented energy bands.

We review the literature in Section 2 and outline our approach and present algorithm details in Section 3. Experimental results are the subject of Section 4 and we conclude in Section 5.

2 Related Work

The object matching in this paper primarily focuses on vehicle instance recognition or fingerprinting. Koller et al. [13] employed a 3D generic vehicle model parameterized by 12 length parameters to instantiate different vehicles. Line segments from the image are matched to the 2D model edge segments obtained by projecting a 3D polyhedral model of the vehicle into the image plane. This method works well when enough image resolution is available.

Feature-based object recognition methods have flourished in recent years. An extensive review of local feature descriptors can be found in [14]. A large body of work is based on the development of corner-like interest point and associated invariant description [8], [15]. The interest point finds distinctive features with precise location, but its descriptor may not be stable under large perspective change. A representative work using local region-like features is the scale-invariant feature transform (SIFT) method [4]. SIFT-like features cannot be extracted reliably in low resolution images. There is a whole body of work on wide baseline matching that deals with quasi-invariant feature-based matching using 2D/3D constraints. In [7], a stable region feature called the Maximally Stable Extremal Regions (MSER) is developed. MSERs are invariant to affine transformation in both image coordinates and intensity. A robust similarity measure is also developed to establish feature correspondences. An improved blob detector is developed in [10], where a robust method is exploited to move across scale space and overlapping regions are allowed. Our regions features adapts this representation.

For object extraction and grouping, Sivic et al. [6] presented a work on grouping object hypotheses in video frames by tracking image patches over long sequences. Affine covariant patches that can be tracked over a large number of frames and move semirigidly over the sequence are grouped into objects. Queries are matched to learned object representations by matching the patch-based multi-view feature groupings. The strength of this approach is that multiple parts of an object could be matched from many different frames. However, the representation and matching may not lead to exact matches but is more suited to similarity searches. Our strategy of object extraction within sequence is motivated by this approach, but we use different feature representation and matching metric. We customize our across sequence alignment and flexible matching components to suit the resolution constraints as well as the goal of exact matching.

For object matching and classification, there has been significant development in part-based approach in recent years. In [15], objects are represented as a flexible constellation of parts. Scale invariant features (parts) are first detected and a probabilistic model is used to represent the appearance, scale, occlusion, and shape (configuration between parts) of the object class. The model parameters are learned using an EM framework and images are classified in a Bayesian manner. Training is required in this approach and object coverage from the detected features is not guaranteed. Another part-based approach is by [5]. In their work, "informative" overlapping parts (fragments) are selected on the basis of maximizing the information delivered by the fragments about the class (faces, cars, etc.) they represent. Offline training has to be conducted in this approach. The representations developed in these works are too coarse for the purpose of object instance matching. Other related work includes [3], [16], [17], [18], [19], etc.

Another representative part-based object (especially vehicular object) detection method is developed in [20]. A vocabulary of distinctive object parts is automatically constructed from a set of training images. Images are then represented using parts from this vocabulary and the spatial configuration between parts is also modeled. Based on this representation, a learning algorithm is used to automatically learn to detect instances of the object class in new images.

Another vehicle identification algorithm is proposed by Ferencz et al. In [2], they used a hyperfeature for object instance matching, where both local object appearance and location saliency are encoded. By modeling the distributions of comparison metrics on the salient patches and applying the mutual information-based feature selection, a compact representation of the features with high saliency can be build from a single example and efficient object identification can be achieved.

3 Proposed Approach

3.1 Overall Approach

Fig. 2 illustrates our overall approach and it consists of four major steps. We briefly summarize the four steps next and more details follow in Sections 3.3, 3.4, 3.5, and 3.6.

3.1.1 Within Sequence Object Mask Generation

In the standard frame-to-frame tracking process, the pixel-by-pixel ownership for the background and foreground cannot be perfectly assigned due to the inadequate background stabilization and subsequent change detection or imperfect background modeling and subtraction. However, reliable object matching requires the distraction from the background be reduced to the minimum. Therefore, we first need to obtain the precise object ownership mask, given an approximate bounding box for the object. We choose region features for the task since they have good coverage property. Blob-like regions for the key frames are extracted after they are aligned with their neighboring frames within the sequence and the blob configuration and appearance are simultaneously compared using an EMD-based metric. Outliers due to background clutter are also rejected and an accurate object mask is generated in the blob matching process.

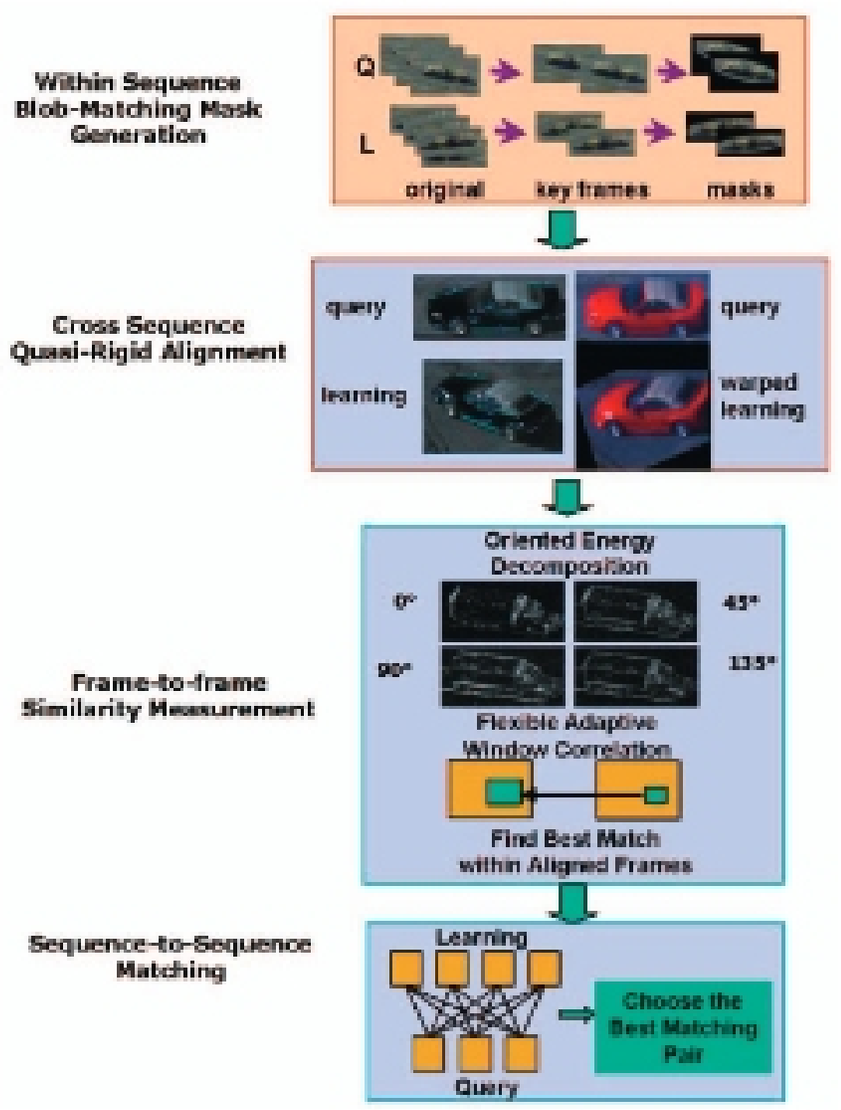

Fig. 2. Overall image matching framework. It consists of *within* sequence mask generation (cream), *across* sequence image alignment (blue) steps, following the similarity measurement. A sequence-to-sequence matching strategy is used to achieve robustness to occlusion, pose, and lighting changes.

3.1.2 Across Sequence Image Alignment and Matching

For across sequence matching between key frames, large pose and appearance change need to be dealt with. Since corner-like and line features have good localization characteristic, they are utilized to align query and learning images to the best possible accuracy.

3.1.3 Matching Measurement

A matching score is produced that consists of several terms (normalized color correlation, color similarity, etc.) that are computed within the object mask. More details are given in Section 3.6.

3.1.4 Sequence-to-Sequence Matching

Finally, we pose the problem of vehicle matching and fingerprinting with the aerial video context as sequence-to-sequence matching problem. Sequence-to-sequence matching can be robust to occlusion, pose, and lighting changes. One frame can potentially find a best match within a sequence of frames that may not be all affected by the same set of changes. We first choose all the representative *key frames* for both learning and query sequences. We then match each key frame in a query sequence to each key frame in a learning sequence. Key frames were selected based on the drastic change in appearance and motion. Simple appearance and motion model were used for this purpose. The frame-to-frame matching uses aggregated matching of local neighborhoods with flexible templates, as illustrated in Section 3.6. The best matching score is chosen as the final match measure. Choosing the best K pairs with/without temporal constraints is another option. Alternatively, one