J. Vis. Commun. Image R. 30 (2015) 208–218

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Robust tracking using visual cue integration for mobile mixed images ☆

Hsiao-Tzu Chen, Chih-Wei Tang *

Department of Communication Engineering, National Central University, Jhongli 32001, Taiwan

ARTICLE INFO

Article history:
Received 21 October 2014
Accepted 15 April 2015
Available online 24 April 2015

Keywords:
Visual tracking
Reflection
Multi-cue integration
Camera motion
Particle filter
Motion compensation
Layer separation
Co-inference

ABSTRACT

The transmitted scene superposed with the reflected scene from a transparent surface leads to mixed images. Few methods have been devoted for tracking on mixed images while such images are ubiquitous in the real world. Thus, this paper proposes a robust single object tracking scheme for mixed images acquired by mobile cameras. Layer separation that decomposes mixed images extracts intrinsic dynamic layers before tracking. In order to make the tracker robust against camera motion, motion compensation is applied to both layer separation and prediction stage of the particle filter. To maximize the observation likelihood and thus optimize particle weights in the face of reflections, the proposed scheme combines sequential importance resampling (SIR) based co-inference and maximum likelihood for multi-cue integration. Experimental results show that the proposed scheme effectively improves tracking accuracy on mixed images with camera motion.

1. Introduction

Visual tracking is essential to many applications of computer vision. Previous trackers tackle problems such as occlusion, illumination variations, pose changes, cluttered background, complex trajectory, and multi-target. But few of them focus on the reflection interference problem. In fact, mixed images contain both reflections from the transparent surface and the transmitted scene behind the surface. Due to being superposed with the reflected scene, the appearance of the target and background in the transmitted scene change significantly in the mixed images and thus inaccurate tracking is easily raised. In complex and dynamic environments, multi-cue integration can improve the robustness of particle filter based trackers. Serby et al. combine multiple low-level features, including interesting points, edges, homogeneous and textured regions, into a particle filter framework for tracking [1]. And the multi-feature observation likelihood is the product of individual likelihoods of different features. To maximize the discriminability of multiple cues, Yang et al. use object detectors to adapt the target observation model [2]. Li et al. propose the weighted Dempter-Shafer fusion to combine evidences from different spatio-temporal SVMs (supper vector machine) into the observation likelihood of a particle filter [3]. The phenomenon of co-inference is proposed by Wu and Huang in 2001 and is refined in [4]. By using the structured variational inference to decouple the dynamics of multiple hidden states, Wu et al. propose co-inference tracking of multiple modalities. The variational parameters of one modality are inferred by the other modalities to maximize the observation likelihood [4]. Sparse representation can be also integrated into the particle filter for data fusion [5–8]. Such trackers are robust against occlusion, illumination variations, and cluttered background. Wu et al. solves a l_1-regularized least squares problem to estimate sparse coefficients of the target candidate [5].

Reflection separation is to estimate the transmitted scene and reflection image from the transparent surface, e.g., glass [9]. Blind source separation aims at estimating the unknown source signals and mixing matrix from a set of mixed signals [10]. For computer vision, reflection separation can take use of blind source separation to estimate source layers. Independent component analysis (ICA) based separation works under assumptions of independent source layers and static mixing, e.g. [11]. Given two mixed images, two source images can be separated by minimizing their structural correlations, e.g. [12]. Under the sparse prior over derivative filters on natural images [13], manually marked edges can help separation from a single image [14]. Automatic separation of weak reflection from a single image can be achieved by using the smoothness constraint and reducing the structural correlation between layers [15]. In [16], several structural priors in the transmitted and reflected layers are combined. Then the geometrical alignment of the reflection region in multiple mixed images is optimized using the augmented Lagrangian multiplier. Gai et al. consider the diversities of layer motions and model the transformation of reflected layers in a parametric way [17].

☆ This paper has been recommended for acceptance by Yehoshua Zeevi.

* Corresponding author.
E-mail addresses: tzuchen248@gmail.com (H.-T. Chen), cwtang@ce.ncu.edu.tw (C.-W. Tang).

http://dx.doi.org/10.1016/j.jvcir.2015.04.006
1047-3203/

Instead, Li et al. assume the parametric transformation for transmitted layers so that variations of reflection layers can be handled [18]. The method in [19] tracks reflection regions in video frames. Another kind of layer separation is to derive intrinsic images including the illumination and reflectance images where the input image is the product of separated layers [13,20]. The intrinsic image, the mid-level description of scenes, is defined by Barrow and Tenenbaum [20]. Sometimes, computer vision algorithms working on such descriptions achieve better performance.

Nowadays, few methods have been proposed for tracking on mixed images. Before using the Kanade–Lucas–Tomasi feature tracker for tracking an object in regions of reflections, the method in [21] applies layer separation [13] to temporally aligned frames to extract the background and foreground layers. Tracks of separated layers are longer than those of the mixed images. Since the focus of tracking is the target but neither the transmitted scene nor the reflected scene, layer separation [13] also extracts the dynamic layer before single object tracking in [22]. Based on the framework of particle filter with compensated motion model [23], the correction stage reweights particles using RGB and [I, R-G, Y-B] color histograms of the mixed images [20]. The [I, R-G, Y-B] color histogram is generated with the aid of a mask, indicating the dynamic regions on the mixed image. Then each particle weight is optimized using maximum likelihood. One problem with Chen et al. [22] is that tracking accuracy will decrease if videos have camera motion. This is mainly because the inaccurate static layer (i.e., reflectance image) makes edges of background and reflections contaminate the dynamic layer (i.e. illumination image). The measurement that refers to the inaccurate mask decreases estimation accuracy of the correction stage. Since few of previous trackers tackle the problem of reflection interference and multi-target tracking should discuss the areas of interactive multiple motion (IMM) model, data association, and state estimation in depth [24], this paper focuses on how to achieve robust single object tracking using multiple cue integration for mobile mixed images. This paper improves the work in [22] and its major contributions are stated as follows. (1) The proposed scheme improves tracking accuracy under the condition of reflections by combining co-inference [4] and maximum likelihood for visual cue integration. (2) The proposed particle filter based scheme realizes co-inference using sequential importance resampling (SIR) [25] instead of sequential important sampling (SIS) to avoid the degeneracy problem. (3) Layer separation with motion compensation for mobile images is proposed to extract objects with active motion. As a result, the proposed scheme significantly improves tracking accuracy on mobile mixed images. The remainder of this paper is organized as follows. Section 2 reviews the compensated motion model for tracking with mobile cameras [23]. Section 3 proposes motion compensated layer separation for images with camera motion. Section 4 proposes a robust single object tracking scheme that combines co-inference [4] and maximum likelihood for mixed images. Section 5 analyzes experimental results and Section 6 concludes this paper.

2. Overview of the compensated motion model for tracking on mobile images

The particle filter (PF) implements the Bayesian filter recursively using the sequential Monte Carlo method [26]. Bayesian tracking consists of the prediction and correction stages to estimate the target state over the posterior probability density function (pdf). Prediction obtains the prior pdf of the target state, $\mathbf{x}_t$, at time t by

$$p(\mathbf{x}_t|\mathbf{z}_{1:t-1}) = \int p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{z}_{1:t-1})d\mathbf{x}_{t-1}, \tag{1}$$

where $\mathbf{z}_{1:t-1} = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{t-1}\}$ is the set of observations up to time t-1. The correction stage updates the posterior pdf $p(\mathbf{x}_t|\mathbf{z}_{1:t})$ using the Bayes' rule by

$$p(\mathbf{x}_t|\mathbf{z}_{1:t}) = \frac{p(\mathbf{z}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{z}_{1:t-1})}{p(\mathbf{z}_t|\mathbf{z}_{1:t-1})}, \tag{2}$$

where $p(\mathbf{z}_t|\mathbf{z}_{1:t-1})$ is the normalization constant, depending on the likelihood function $p(\mathbf{z}_t|\mathbf{x}_t)$. For particle filter, the posterior pdf is approximated by a random measure, $\{\mathbf{x}_{1:t}^{(i)}, \pi_t^{(i)}\}_{i=1}^N$, where $\{\mathbf{x}_{1:t}^{(i)}, i = 1, \ldots, k\}$ is a set of particles with the associated importance weights $\{\pi_t^{(i)}, i = 1, \ldots, N\}$. The samples $\mathbf{x}_t^{(i)}$ are drawn from an importance density $q(\mathbf{x}_t^i|\mathbf{x}_{1:t-1}^i, \mathbf{z}_{1:t})$ and their weights are updated by

$$\pi_t^{(i)} = \pi_{t-1}^{(i)} \frac{p(\mathbf{z}_t|\mathbf{x}_t^{(i)})p(\mathbf{x}_t^{(i)}|\mathbf{x}_{t-1}^{(i)})}{q(\mathbf{x}_t^{(i)}|\mathbf{x}_{1:t-1}^{(i)}, \mathbf{z}_{1:t})}, \tag{3}$$

where $q(\mathbf{x}_t|\mathbf{x}_{1:t-1}^{(i)}, \mathbf{z}_{1:t})$ chose to be $p(\mathbf{x}_t|\mathbf{x}_{t-1}^{(i)})$ can reduce the degeneracy problem of sequential importance sampling [26].

In visual tracking, both object motion and camera motion should be handled. The compensated motion model that includes the control vector of camera motion can improve the tracking accuracy significantly [23]. In [23], the state vector $\mathbf{x}_t^i$ of the ith particle at time t is predicted by

$$\mathbf{x}_t^{(i)} = \begin{bmatrix} S_{x,t}^{(i)} \\ S_{y,t}^{(i)} \\ W_{x,t}^{(i)} \\ W_{y,t}^{(i)} \\ H_{x,t}^{(i)} \\ H_{y,t}^{(i)} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} S_{x,t-1}^{(i)} \\ S_{y,t-1}^{(i)} \\ W_{x,t-1}^{(i)} \\ W_{y,t-1}^{(i)} \\ H_{x,t-1}^{(i)} \\ H_{y,t-1}^{(i)} \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} G_{x,t}^{(i)} \\ G_{y,t}^{(i)} \end{bmatrix} + \varepsilon_t^{(i)}, \tag{4}$$

where $[S_{x,t}^{(i)}, S_{y,t}^{(i)}]^T$ and $[H_{x,t-1}^{(i)}, H_{y,t-1}^{(i)}]^T$ are the position and scale of the target, respectively, $\varepsilon_t^{(i)}$ is Gaussian noise [27], $[G_{x,t}, G_{y,t}]^T$ is camera motion, and $[W_{x,t}^{(i)}, W_{y,t}^{(i)}]^T$ is the object motion on the 2-D image after motion compensation.

3. Layer separation using motion compensation for mobile images

An image can be represented as the product of the reflectance and illumination images. By layer separation, an image can be decomposed into intrinsic images, including a reflectance image and an illumination image [13,20]. This is true no matter whether an image has reflections from the transparent surface. Assume that the reflectance is constant and the illumination changes. Layer separation in [13] estimates the intrinsic images based on the sparse prior over derivative filters on the previous T frames. For videos, the estimated reflectance and illumination images correspond to the static and dynamic layers of images, respectively. That is, moving objects will be separated into the dynamic layer if the video has camera motion. For tracking on mixed images with camera motion, the pre-processing stage expects to extract the dynamic layer that contains objects with active but not passive (camera) motion most of the time. Thus, instead of using reflection separation to get the transmitted scene and reflection image from the transparent surfaces, we integrate the layer separation method in [13] with motion compensation to estimate dynamic layers of mixed images, i.e., illumination images. Motion information from such layers significantly improves tracking accuracy on mixed images.

For intrinsic images in the log domain [13],

$$i(x, y, t) = l(x, y, t) + r(x, y), \tag{5}$$