JOINT TRACKING AND VIDEO REGISTRATION BY FACTORIAL HIDDEN MARKOV MODELS

Xue Mei and Fatih Porikli[†]

University of Maryland
College Park, MD 20742
xuemei@umiacs.umd.edu

[†]Mitsubishi Electric Research Labs
Cambridge, MA 02139
fatih@merl.com

ABSTRACT

Tracking moving objects from image sequences obtained by a moving camera is a difficult problem since there exists apparent motion of the static background. It becomes more difficult when the camera motion between the consecutive frames is very large. Traditionally, registration is applied before tracking to compensate for the camera motion using parametric motion models. At the same time, the tracking result highly depends on the performance of registration. This raises problems when there are big moving objects in the scene and the registration algorithm is prone to fail, since the tracker easily drifts away when poor registration results occur. In this paper, we tackle this problem by registering the frames and tracking the moving objects simultaneously within the factorial Hidden Markov Model framework using particle filters. Under this framework, tracking and registration are not working separately, but mutually benefit each other by interacting. Particles are drawn to provide the candidate geometric transformation parameters and moving object parameters. Background is registered according to the geometric transformation parameters by maximizing a joint gradient function. A state-of-the-art covariance tracker is used to track the moving object. The tracking score is obtained by incorporating both background and foreground information. By using knowledge of the position of the moving objects, we avoid blindly registering the image pairs without taking the moving object regions into account. We apply our algorithm to moving object tracking on numerous image sequences with camera motion and show the robustness and effectiveness of our method.

Index Terms— Tracking, video registration, factorial Hidden Markov Model, camera motion

1. INTRODUCTION

Visual tracking is a critical task in many computer vision applications such as surveillance, robotics, human computer interaction, vehicle tracking and medical imaging, etc. Tracking works by finding a region in the current frame that matches a template as closely as possible. In the following paragraphs, we briefly review the conventional tracking and image registration methods.

Mean shift [1] is a nonparametric density gradient estimator to find the image window that is most similar to the object's color histogram in the current frame. It iteratively carries out a kernel based search starting at the previous location of the object. The success of the mean shift highly depends on the discriminating power of the histograms that are considered as the objects' probability density function.

Tracking can be considered as an estimation of the state for a time series state space model. The problem is formulated in probabilistic terms. Early works use a Kalman filter to provide solutions that are optimal for a linear Gaussian model. The particle filter, also known as the sequential Monte Carlo method [2], is the most popular approach. It recursively constructs the posterior pdf of the state space using Monte Carlo integration. It has been developed in the computer vision community and applied to tracking problems under the name Condensation [3]. Previously, subspace representations were successfully used for tracking by finding the minimum distance from the tracked object to the subspace spanned by the training data or previous tracking results [4, 5]. Particle filter is based on random sampling that becomes a problematic issue due to sample degeneracy and impoverishment. Tracker is prone to fail due to the contamination of the model subspace.

Tracking can also be considered as a classification problem and a classifier can be trained to distinguish the object from the background [6]. This is done by constructing a feature vector for every pixel in the reference image and training a classifier to separate pixels that belong to the object from pixels that belong to the background. One obvious drawback of the local search methods is that they tend to get stuck into the local optimum.

A covariance based object description that fuses different types of features and modalities is used to successfully track nonrigid objects [7]. The shortcoming of the algorithm is the exhaustive search in the local area for the candidate location and hard to handle large camera motion.

Image registration amounts to establishing a common frame of reference for a set of images of the same scene taken at different times, from different views, or by different sensors. It plays a vital role in many computer vision applications such as video tracking, medical imaging, remote sensing, super-resolution and data fusion. Several comprehensive surveys [8, 9] on image registration have been published to cover the progress achieved in this rich area.

SIFT features are used to register images in an approach that is insensitive to the ordering, orientation, scale and illumination of the images and that removes the 'outlier' image, which does not have any overlapping area with the other images [10]. Mutual information based registration [11, 12], which is inspired by information theory, is considered among the state-of-the-art registration methods for multi-modality images [13]. The assumption that the intensities between corresponding pixels are similar no longer holds, while the distribution of the intensities of matched pixels should be maximally dependent. To improve the convergence properties, [14] applied global estimation on the common information. In [15], it proposed to register the images by iteratively minimizing the orientation distance of high intensity gradient pixels using second and third order spatial gradients.

Tracking of independently moving objects in the case of a mov-

ing camera is inherently more challenging and complex, since the motion of the camera induces a motion in all pixels in the image. [16] represented and modeled the scene in terms of a small group of motions. By incorporating spatial constraints and given assumptions about the expected level of model failure, [17] estimated the number of motion models automatically. The tracking result highly depends on the quality of the registration which is unreliable when the registration algorithm fails to achieve reasonable results.

To overcome the shortcomings of existing approaches, we propose an algorithm based on factorial Hidden Markov Model framework [18] which is successfully used in [19] to handle occlusions during tracking. Under this framework, tracking and registration work jointly. We dub our joint registration and tracking algorithm JTR in the rest paper. Background and foreground information is incorporated in the framework and the tracking and registration results are reciprocal. The tracking score is obtained by incorporating both background and foreground information. By using knowledge of the position of the moving objects, we avoid blindly registering the image pairs without taking the moving object regions into account.

The rest of the paper is organized as follows. In the next section, the probabilistic joint framework and overview of the algorithm are presented. In Section 3, a fast image registration algorithm is proposed. Section 4 describes the covariance tracker and our improvements on it. Our results on image sequences are presented and discussed in Section 5. In Section 6, we conclude our work.

2. JOINT FRAMEWORK

In this section, we describe the joint framework built on the framework of factorial Hidden Markov Models (HMM) [18] and show the advantages we can get by applying this framework. In an HMM, the past information is conveyed through the single hidden state. Factorial HMM generalizes it by factorizing the hidden state into multiple state variables and is therefore able to handle more complex problems described by this graphical model. In stead of only factorizing the hidden state, we factorize both the observation and hidden state. We present a structured approximation to yield a tractable algorithm and infer the parameters by decoupling both the observation and state variables.

The probabilistic factorial Hidden Markov Model framework consists of two parts: the state we are going to infer and the observation from the image sequences. In our framework, the state variable X_t is decomposed to camera motion (registration) parameters X_t^c and moving object motion (tracking) parameters X_t^o, where X_t denotes the state X at time t. By incorporating the registration and tracking in the same framework, they mutually benefit each other by interacting. The tracking and registration task is to infer $X_t = (X_t^o, X_t^c)$ based on all the observed image evidence $\underline{Z}_t = \{Z_1, Z_2, \cdots, Z_t\}$, where $Z_t = (Z_t^o, Z_t^b)$is the image observation at time t. We further factorize the observation Z_t to two parts: moving object observation Z_t^o and background observation Z_t^b.

The tracking process is viewed as density propagation [4] from $p(X_{t-1}|\underline{Z}_{t-1})$ to $p(X_t|\underline{Z}_t)$ and the propagation equation is given by

$$p(X_t|\underline{Z}_t) \propto p(Z_t|X_t) \int p(X_t|X_{t-1}) p(X_{t-1}|\underline{Z}_{t-1}) dX_{t-1} \quad (1)$$

In addition, since the camera motion and moving object motion are independent, we have

$$\begin{aligned} p(X_t|X_{t-1}) &= p(X_t^o, X_t^c|X_{t-1}^o, X_{t-1}^c) \\ &= p(X_t^o|X_{t-1}^o) p(X_t^c|X_{t-1}^c) \end{aligned} \quad (2)$$

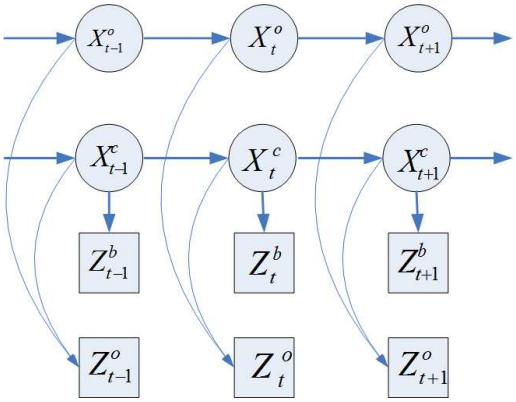

Fig. 1. A directed acyclic graph (DAG) specifying conditional independence relations for the factorial Hidden Markov Model of the tracking and registration framework.

What's more, since the background observation is independent of object motion, the observation probability given the state is derived as

$$\begin{aligned} p(Z_t|X_t) &= p(Z_t^o, Z_t^b|X_t^o, X_t^c) \\ &= p(Z_t^o|Z_t^b, X_t^o, X_t^c) p(Z_t^b|X_t^o, X_t^c) \\ &= p(Z_t^o|X_t^o, X_t^c) p(Z_t^b|X_t^c) \end{aligned} \quad (3)$$

Substituting Eqn.2 and Eqn.3 into Eqn.1, we obtain

$$\begin{aligned} p(X_t|\underline{Z}_t) \propto p(Z_t^o|X_t^o, X_t^c) p(Z_t^b|X_t^c) \int p(X_t^o|X_{t-1}^o) \\ p(X_t^c|X_{t-1}^c) p(X_{t-1}|\underline{Z}_{t-1}) dX_{t-1} \end{aligned} \quad (4)$$

Figure 1 illustrates the conditional independence relations for the Hidden Markov Model of the tracking and registration framework.

2.1. Algorithm Overview

Assuming at the first two frames, we have the registration result X_0^c and tracking result X_0^o. We manually select the moving objects or detect them using a detection algorithm in the first few frames. The registration result is obtained by registering the first two frames using the method we described next after taking out the moving objects in the images. Then, for the rest frames I_i, where $i = 3, 4, \cdots$, JTR is summarized in Algorithm 1.

3. FAST IMAGE REGISTRATION VIA JOINT GRADIENT MAXIMIZATION

We present a new, fast and efficient method for image registration. Our approach applies a novel similarity function on the image locations which have high gradient magnitudes. The similarity function secures a fast convergence. The parametric motion transformation parameters p is obtained by maximizing the joint gradient function. It is defined as

$$p^* = arg \max_p \sum_{(x_i, y_i) \in S} (E_1(x_i, y_i, p) + E_2(x_i, y_i))^2 \quad (5)$$

where E_1 and E_2 represent the edge (or energy) images of I_1 and I_2 which are generated by applying Canny edge detector. Edges are

Algorithm 1 Joint Tracking and Registration

1: Particles are drawn from a Gaussian distribution which has mean X_{i-1}^c. Registration scores are obtained by registering the whole images I_i and I_{i-1} according to the registration particle parameters.
2: The registration scores are ranked from highest to lowest and the top 10 particles which have the highest registration scores are picked out. We name them as $p(Z_i^{bj}|X_i^{cj})$, where $j = 1, 2, \cdots, 10$.
3: Particles are drawn from a Gaussian distribution which has mean X_{i-1}^o. The registration scores are recalculated by removing the candidate moving object in the image according to the moving object particle parameters. These probabilities are $p(Z_i^{bjk}|X_i^{cjk})$ where $k = 1, 2, \cdots, m$, and m is the number of particles we draw as the moving object parameters.
4: Tracking scores are obtained from the tracker and the probabilities are $p(Z_i^{ojk}|X_i^{ojk}, X_i^{cjk})$.
5: The registration and tracking scores are substituted to Eqn.4 and get the combination score of both registration and tracking.
6: The probability $p(X_i^{jk}|\underline{Z}_i^{jk})$ which achieves the highest score is the registration and tracking result.

the locations where image has depth discontinuities and high information values. Applying maximization to a small set of salient edge pixels S makes our method fast and robust. Detailed descriptions can be found in [20].

4. PARTICLE TRACKING WITH COVARIANCE FEATURES

In [7], it proposes an algorithm which uses covariance matrix to model the appearance of the object. At each frame, a feature image is constructed. For a given object region, the covariance matrix of the features as the model of the object is computed. In the current frame, the candidate regions are cropped out according to the transformation parameters drawn with respect to the state distribution. We find the region that has the minimum covariance distance from the model as the tracking result.

We improve their algorithm by using particle filter to draw object motion parameters that denote candidate object position. In stead of applying exhaustive search in the local area for the moving object position, we draw particles for the candidate positions according to a normal distribution. At each frame, we take the region cropped out according to the sample drawn from the distribution which has the smallest distance from the current object model. The best matching region determines the location of the object in the current frame.

The probability of the tracking result given the camera motion parameters and moving object parameters is written as

$$p(Z_t^o|X_t^o, X_t^c) = exp\{-\rho\} \tag{6}$$

where ρ is the similarity function used to calculate the distance between the covariance matrix of the rectangular region.

5. EXPERIMENTAL RESULTS

We have conducted a through analysis and tested our algorithm with a large number of moving camera sequences including large moving objects. JTR was tested on both visible and IR image sequences and performs well in terms of following the object position.

The geometric transformation parameters between consecutive frames are modeled by affine transformation parameters $\vec{p} = (p_1, p_2, p_3, p_4, p_5, p_6)^T$. The model is given by

$$x_i^2 = p_1 x_i^1 + p_2 y_i^1 + p_3, \qquad y_i^2 = p_4 x_i^1 + p_5 y_i^1 + p_6 \tag{7}$$

where x_i^1, y_i^1 and x_i^2, y_i^2 are the pixel coordinates before and after transformation, respectively. The tracking area is described by a rectangular window modeled by a 2-dimensional state vector $X_0^o = [x_0, y_0]$, where (x_0, y_0) represents the centroid of the tracking window. The width and height of the tracking window are fixed for the tracking sequences. Currently the parameters are initialized manually.

In one of the experiments, an infrared image sequences consisting of 320×256 color videos, recorded at 30 frames/second were used. The target-to-background contrast is very low and the noise level is high for the IR images. Figure 2 shows the results of JTR (left column) and the covariance tracker without applying registration (right column). The second tracker can't catch up to the movement of the moving object on the top because the camera and the moving object is moving in the opposite direction and there is a big motion between consecutive frames. With applying registration in the tracking, JTR is effective under low contrast and noisy situations in tracking both the moving objects in the image sequences.

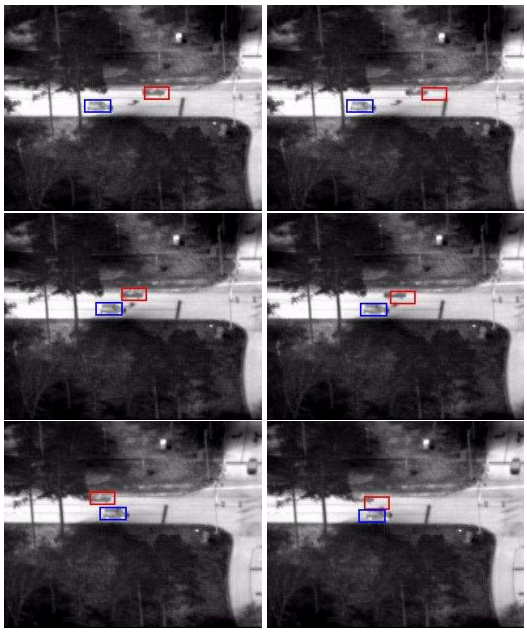

Fig. 2. Left column: Tracking results of JTR. Right column: Tracking results of the covariance tracker without applying registration.

Figure 3 shows the additional tracking results by applying JTR. There is big moving object in the scene with big camera motion between consecutive frames. JTR is very robust against the big object motion and camera motion.

Table 1 shows tracking results for JTR and the covariance tracker after registration without taking moving objects into account during registration (comparison method). In the table, MO stands for moving object and CM stands for camera motion. The percentage is the ratio of the number of frames in which the moving objects are successfully being tracked to the total number of frames. In the small MO/small CM scenario, both methods work pretty well. While in

Fig. 3. More tracking results for JTR.

	estimation errors	
scenario	comparison method	JTR
small MO/small CM	95.90%	95.48%
small MO/large CM	73.96%	91.36%
large MO/small CM	82.01%	88.19%
large MO/large CM	34.54%	81.72%

Table 1. Tracking results for JTR and the covariance tracker after registration without taking moving objects into account during registration (comparison method). In the table, MO stands for moving object and CM stands for camera motion.

the small MO/large CM and large MO/small CM scenarios, JTR works better than the comparison method to some extent. In the large MO/large CM scenarios, the performance of the comparison method drops sharply due to its limitation in handling large object moving and large camera movement at the same time. On the contrary, JTR achieves reasonable results in this scenario.

6. CONCLUSIONS

In this paper, we have proposed an efficient and robust simultaneous registration and tracking algorithm using a factorial Hidden Markov Model. We have demonstrated our algorithm for tracking moving objects in various challenging sequences. By registering background, we compensate for the motion of the camera movement and the registration and tracker work together to achieve robust tracking results. We propose a framework which can handle registration and tracking at the same time. Under this framework, tracking and registration are not working separately, but mutually benefit each other by interacting. We improve covariance tracker by using particle filter. It avoids exhaustive searching in the local area which is prone to fail in the large motion induced by the camera movements.

7. REFERENCES

[1] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, pp. 564–577, 2003.

[2] A. Doucet, N. de Freitas, and N. Gordon, Eds., *Sequential Monte Carlo Methods in Practice*, Springer-Verlag, New York, 2001.

[3] M. Isard and A. Blake, "Condensation - conditional density propagation for visual tracking," *Int. J. Computer Vision*, vol. 29, pp. 5–28, 1998.

[4] M. J. Black and A. D. Jepson, "Eigentracking: Robust matching and tracking of articulated objects using a view-based representation," *Int. J. Computer Vision*, vol. 26, pp. 63–84, 1998.

[5] M.-H. Y. J. Ho, K.-C. Lee, and D. Kriegman, "Visual tracking using learned subspaces," *Proc. IEEE Conf. on Comp. Vision and Patt. Recog.*, pp. 782–789, 2004.

[6] S. Avidan, "Ensemble tracking," *Proc. IEEE Conf. on Comp. Vision and Patt. Recog.*, pp. 494 – 501, 2005.

[7] F. Porikli, O. Tuzel, and P. Meer, "Covariance tracking using model update based on lie algebra," *Proc. IEEE Conf. on Comp. Vision and Patt. Recog.*, pp. 728–735, 2006.

[8] L. Brown, "A survey of image registration techniques," *ACM Computing Surveys*, vol. 24, pp. 325–376, 1992.

[9] R. Szeliski, "Image alignment and stitching: a tutorial," *Microsoft Research MSR-TR-2004-92*, 2004.

[10] M. Brown and D. G. Lowe, "Recognising panoramas," *IEEE International Conference on Computer Vision*, pp. 1218–1225, 2003.

[11] A. Collignon, F. Maes, D. Vandermeulen, P. Suetens, and G. Marchal, "Automated multimodality image registration using information theory," *Information Processing in Medical Imaging*, pp. 263–274, 1995.

[12] P. Viola and W. M. WellsIII, "Alignment by maximization of mutual information," *International Conference on Computer Vision*, pp. 16–23, 1995.

[13] J. Pluim, J. Maintz, and M. Viergever, "Mutual information based registration of medical images: a survey," *IEEE Trans. on Medical Imaging*, vol. 8, pp. 986–1004, 2003.

[14] M. Irani and P. Anandan, "Robust multi-sensor image alignment," *IEEE International Conference on Computer Vision*, pp. 959–966, 1998.

[15] Y. Keller and A. Averbuch, "Multisensor image registration via implicit similarity," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, pp. 794–801, 2006.

[16] S. Ayer and H. S. Sawhney, "Layered representation of motion video using robust maximum-likelihood estimation of mixture models and mdl encoding," *International Conference on Computer Vision*, pp. 777–784, 1995.

[17] Y. Weiss and E. H. Adelson, "A unified mixture framework for motion segmentation: incorporating spatial coherence and estimating the number of models," *Proc. IEEE Conf. on Comp. Vision and Patt. Recog.*, pp. 321–326, 1996.

[18] Z. Ghahramani and M. Jordan, "Factorial hidden markov models," *Machine Learning*, vol. 29, pp. 245–275, 1997.

[19] Y. Wu, T. Yu, and G. Hua, "Tracking appearances with occlusions," *Proc. IEEE Conf. on Comp. Vision and Patt. Recog.*, vol. 1, pp. 789–795, 2003.

[20] X. Mei and F. Porikli, "Fast image registration via joint gradient maximization: application to multi-modal data," *Proceedings of SPIE Volume 6395 Electro-Optical and Infrared Systems: Technology and Applications III*, 2006.