Visual Tracking using the Joint Inference of Target State and Segment-based Appearance Models

Junha Roh, Dong Woo Park, Junseok Kwon and Kyoung Mu Lee
Department of Electrical and Computer Engineering, Seoul National University, Seoul, Korea.
rohjunha@gmail.com, kospi1981@gmail.com, paradis0@snu.ac.kr, kyoungmu@snu.ac.kr

***Abstract*—In this paper, a robust visual tracking method is proposed by casting tracking as an estimation problem of the joint space of non-rigid appearance model and state. Conventional trackers which use templates as the appearance model do not handle ambiguous samples effectively. On the other hand, trackers that use non-rigid appearance models have low discriminative power and lack methods for restoring methods from inaccurately labeled data. To address this problem, multiple non-rigid appearance models are proposed. The probabilities from these models are effectively marginalized by using the particle Markov chain Monte Carlo framework which provides an exact and efficient approximation of the joint density through marginalization and the theoretical evidences of convergence. An appearance model combines multiple classification results with different features and multiple models can infer an accurate solution despite the failure of several models. The proposed method exhibits high accuracy compared with nine other state-of-the-art trackers in various sequences and the result was analyzed both analyzed both qualitatively and quantitatively.**

I. Introduction

Visual tracking is one of the most important application in the field of computer vision and a few recent trackers have shown promising results [1], [2], [3]. Most existing trackers use templates or patches as appearance models through bounding-box representation for its simplicity and regularity. However, these models cannot effectively distinguish between the target and its background, particularly for deformable objects. These models may also learn ambiguous templates which misalign the target.

Several non-rigid appearance models have been employed to address this problem [4], [5], [6], [7], [8]. Appearance models that use elements with non-regular shapes, such as pixels or superpixels, can separate the target precisely from its background. However, these models have low discriminative power and are prone to misclassifying elements.

In [4], the authors trained a random forest to force each pixel to vote for a center position of the target. The center position is then estimated from the votes. The pixels voted to the estimated center are regarded as pixels with positive labels. This method can sometimes work well with highly deformed objects, but it often works poorly. Failure occurs when several pixels in the background vote for the estimated center or when several pixels in the target vote for the wrong position by mistake or because of ambiguity. Once wrongly labeled pixels are included in the positive pixels, the results worsen because of the lack of a reintegration step.

Another approach was introduced in [5], where elements are clustered by color and their confidence scores of being target elements are computed. Although this approach is a simple and effective means of classifying elements, it still exhibits several problems. First, spatial information is not considered effectively. All cluster index assignments are computed in the feature space and not in the spatial domain. For example, a red color may be the discriminative feature in several parts of the target but not in other parts. Therefore, simply clustering features without spatial information will be hazardous. Second, the approach is highly sensitive to color information. If a target has colors similar to the background, then several clusters may have differently labeled superpixels and may cause an inaccurate estimation of confidence. In this case, other information should be employed to distinguish the target from the background.

To alleviate these problems, we propose a visual tracking method that uses multiple non-rigid appearance models to handle ambiguities near the boundaries of the target object and to combine multiple features effectively. The method casts the tracking problem as an inference problem in the joint space of non-rigid appearance model and state (JISAT). An appearance model linearly combines locally trained classifiers with different features and works as a particle in sequential Monte Carlo (SMC). The proposed method uses the particle Markov chain Monte Carlo (PMCMC) to draw particles and candidate states from the joint space, thus allowing us to obtain the appearance models and a state simultaneously. This method with the multiple feature combination from the multiple appearance models allows us to handle the ambiguity in classification. A number of models fail to classify several elements, but the remaining models with higher weights can correctly classify those elements, thus allowing an accurate estimation of the state possible.

II. Particle Markov Chain Monte Carlo

The Markov chain Monte Carlo (MCMC) is a sampling method that accepts samples with a probability of acceptance ratio. In the Metropolis-Hastings (MH) method, the acceptance ratio is designed as follows:

$$\mathcal{A}(x, x^*) = 1 \wedge \frac{p(x^*)q(x|x^*)}{p(x)q(x^*|x)}, \tag{1}$$

where $\wedge$ is a minimum function, and $q(\cdot|\cdot)$ is a proposal distribution. The main problem of the MCMC is that the

performance is sensitive to the selection of the proposal distribution. The MCMC also exhibits degeneracy of quality from the joint distribution. To overcome these problems, reference [9] proposed the particle marginal MH (PMMH) sampler to update a static variable, θ, and the hidden states, $x_{0:t}$, simultaneously. The key concept of [9] is using an approximate distribution $p_\theta(x_{0:t}|y_{0:t})$ with SMC as a proposal density $q(\cdot|\cdot)$. In each iteration, a static variable θ^* is drawn from $q(\cdot|\theta)$, and an SMC draws the samples $x^*_{0:t}$ from an estimated distribution $\hat{p}_{\theta^*}(\cdot|y_{0:t})$. The marginal likelihood estimate $\hat{p}_{\theta^*}(y_{0:t})$ and acceptance ratio are then computed as follows:

$$1 \wedge \frac{\hat{p}_{\theta^*}(y_{0:t})p(\theta^*)}{\hat{p}_\theta(y_{0:t})p(\theta)} \frac{q(\theta|\theta^*)}{q(\theta^*|\theta)}. \quad (2)$$

The marginal density instead of the joint density, is notably involved in the expression. Therefore, we only design $q(\theta^*|\theta)$ and convergence is guaranteed under a mild assumption, as detailed in [9].

III. Tracking via Sampling Appearance Models-State Space

The details of the proposed method are presented in this section.

A. Non-rigid Appearance Model

The proposed method employs boosted decision trees as classifiers and superpixels [10] as elements for representation. An image is divided into a grid of overlapped patches, P^i, to use the spatial information. For each local region, a set of superpixels, $\{S^i\}$, is constructed, and F classifiers, $C^{i,f}$, are allocated where each classifier uses a different feature. The N appearance models, $M^{i,j} = (\{w^{i,j,f}\}, \theta^{i,j})$, are also allocated to a region. A model or a particle combines the classification results using weights, $\{w^{i,j,f}\}$, and then thresholds using a certain value, $\theta^{i,j}$. Each particle represents a different combination and confidence of the classification results. The particles generate a binary patch and color histogram for evaluation according to the probability that

$$p(M^{i,j}_t) = p(\theta^{i,j}) \prod_f p(w^{i,j,f}), \quad (3)$$

where $\theta^{i,j} \sim \mathcal{N}(0.5, \sigma_{th})$, $w^{i,j,f} \sim \mathcal{N}(w^{i,j,f}_0, \sigma_w)$, and $w^{i,j,f}_0$ is an initial weight. These values are also propagated along the SMC steps through $\mathcal{N}(0, \sigma_w)$ and $\mathcal{N}(0, \sigma_{th})$. The role of a single particle is illustrated in Fig. 1.

B. Sampling Joint Variables

1) Appearance Model Construction: In this method, a tracking problem is casted as an estimation of the maximum a posteriori (MAP) solution of the joint space of the appearance models and a state through PMMH. First, the appearance models are drawn to provide a simple proposal distribution. Second, the states are sampled using approximated marginal distributions as computed earlier. Most of the procedures in SMC, the first step, have been described except for the particle evaluation and marginalization of binary patches. Particle evaluation is conducted by comparing the histograms of the particles and a model. We assume that the particles with high weights represent the target efficiently. We then define a model with K particles with the highest weights. We evaluate the particles by comparing the histograms with the model and computing the overlapped regions. The binary masks and histograms can be compared with a particle because the model consists of particles. The weight of the j-th particle from the m-th model component is computed by

$$w^{i,j,m} = \exp\left(\lambda \times \frac{|r^{i,j} \cap R^{i,m}|}{|r^{i,j} \cup R^{i,m}|} \times S(h^{i,j}, H^{i,m})\right), \quad (4)$$

where $r^{i,j}$ and $R^{i,m}$ are the regions of the particle and the model, respectively; and $h^{i,j}$ and $H^{i,m}$ are their respective histograms. The final weight of the particle is then determined by

$$w^{i,j} = \sum_m W^{i,m} w^{i,j,m}. \quad (5)$$

The marginal probability of a superpixel can be computed from the particles by

$$p(S_k) = \sum_{i,j} w^{i,j} l^{i,j}_k, \quad (6)$$

where S_k denotes the k-th superpixel and $l^{i,j}_k$ denotes the label of the k-th superpixel determined by $M^{i,j}$. The superpixel index is notably unrelated to the local region index i because the superpixel can lie on multiple regions. The marginalization of the local responses is illustrated in Fig. 2.

2) Sampling States: When the appearance models are drawn, the MH method is used to draw the state candidates. A state candidate, X^*, is simply drawn from the multivariate Gaussian distribution with a fixed σ and $X^{(i)}$ as a mean. The sample is accepted with the probability of the acceptance

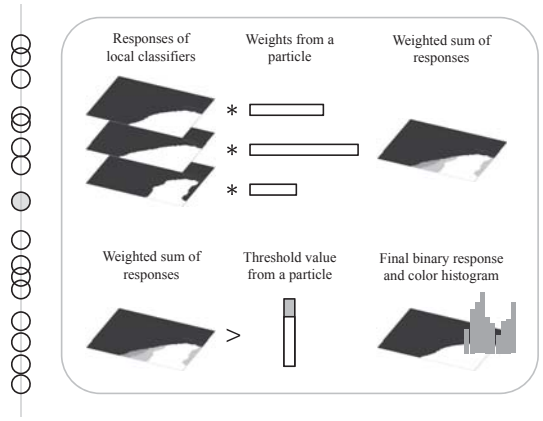

Fig. 1. A single particle representing an appearance model combined with the responses of the local classifiers. The weights of the responses and a thresholding value generate a binary classification response. A color histogram is then extracted from the mask.

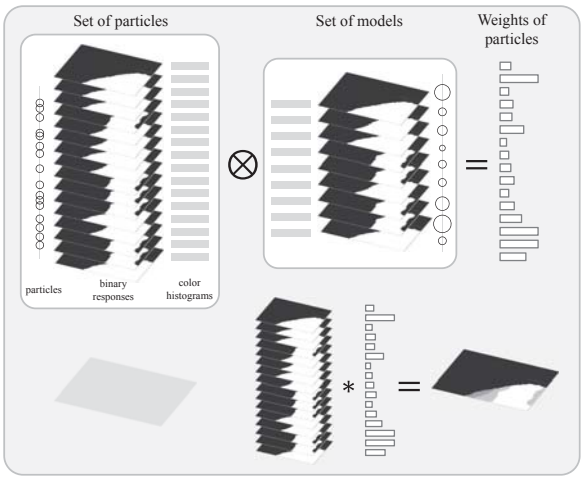

Fig. 2. Marginalized local responses. A local patch has N particles, which indicate multiple binary segmentations and color histograms. The responses must be marginalized to compute the acceptance ratio in the MH step. The first row shows the computation of the particle weights using histograms and masks. The second row shows the marginalization using binary segmentation masks and the weights computed earlier. K particles with the largest weights are chosen to construct the model.

ratio after sampling. To compute the acceptance ratio, $p_{X^*}(Y)$, which is a marginal probability given a candidate X^*, must be computed by

$$\hat{p}_{X^*}(Y) = \sum_{(x,y)\in R^*} p(x,y) = \sum_{(x,y)\in R^*} \sum_k \mathbb{I}_{S_k}(x,y) * p(S_k), \tag{7}$$

where (x, y) is the pixel position in the image, R^* is the region of X^*, and $\mathbb{I}_{S_k}(\cdot,\cdot)$ is an indicator that returns 1 if the position is within the region of a superpixel S_k. The acceptance ratio is computed by

$$\mathcal{A}(X^{(i)}, X^*) = 1 \wedge \frac{\hat{p}_{X^*}(Y)}{\hat{p}_{X^{(i)}}(Y)} \frac{q(X^{(i)}|X^*)}{q(X^*|X^{(i)})}. \tag{8}$$

3) Model Update: The PMMH jointly draws samples as described in Eqs. (3-8) for each frame t in the sequence and stores a joint variable $(X_t^{(i)}, \left\{M_t^{i,j}\right\})$ with the highest probability. $X_t^{(i)}$ then becomes the estimated state in t, and $\left\{M_t^{i,j}\right\}$ are used as the initial particles for the next frame. Drawing all particles in each iteration is impractical. Therefore, the particles are drawn once in each frame and are then reused, which does not harm the convergence of PMMH as stated in [9]. To update the model, the best K particles among the stored particles are collected as the model for each N^t frame. Similarly, the local classifiers are also retained using the collected superpixels with labels. The labels of the superpixels are generated by thresholding the local classifier responses with 0.5.

IV. Experimental Results

For the experiments, seven video sequences were tested, namely, *basketball*, *diving*, *fx*, *gymnastics*, *faceocc*, *twinings* and *lazysong*. The proposed methods (JISAT-f, -sc, -sz+a, -sc+a) were compared with nine state-of-the-art tracking methods, namely, BHT [11], HT [4], LGT [6], IVT [2], MIL [1], BHMC [7], VTD [3], MTT [12], SPT [5]. The proposed method is implemented using C++, and OpenCV, and tested on a PC with a 3.30 GHz CPU. The superpixel segmentation was obtained by [10] and [13], and the optical flow was computed by [14]. Three kinds of features are used: FREAK [15], optical flow, and color.

Fig. 3. The tracking results of JISAT (with fixed size) in the *basketball* sequence. The frame numbers of the figures in the first row are 5, 25, 45, 65, and 85. The numbers in the second row are 105, 125, 145, 165, and 185.

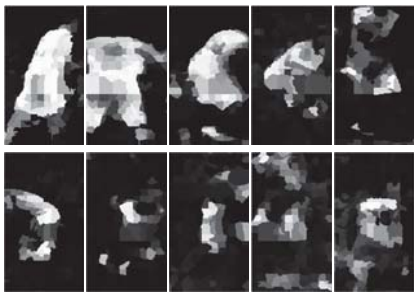

Fig. 4. Marginalized probabilities of the superpixels of JISAT (with fixed size) in the *basketball* sequence. The frame numbers of the figures in the first row are 5, 25, 45, 65, and 85. The numbers in the second row are 105, 125, 145, 165, and 185.

A. Qualitative Analysis

The tracking results and marginal masks of JISAT in a *basketball* sequence are respectively shown in Figs. 3 and 4. The proposed method successfully tracked the target in the *basketball* sequence. In this test sequence, the basketball player moves abruptly and deforms severely even with motion blur and illumination changes. Tracking the target using color information appears easy because the target wears a white uniform. However, the reflections and shadows on the uniform of the player make the color feature unreliable. The uniform also lacks features because of the homogeneity in color. Another challenge in the sequence is the appearance of the other player wearing a uniform with a similar color in #147 to #190. JISAT tracked the target successfully despite these challenges. SPT, which has a structure similar to that of our method, failed to track the target in #125 to #130, as shown in Fig. 5, but recovered the correct target state at #247 because of its false detection of occlusion. HT, another tracker based

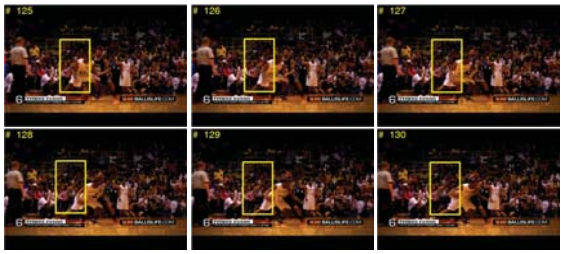

Fig. 5. Failure case of SPT in *basketball* sequence. Frame numbers of figures are 125-130. Due to abrupt motion and static background, the tracker failed to track the target.

Fig. 6. Tracking results of HT in *basketball* sequence. Frame numbers of figures are 4, 5, and 6. Many trackers based on a non-rigid appearance model suffer from the drift problem.

on the non-rigid appearance model, also failed to track the target as illustrated in Fig. 6. Most of the trackers based on rigid appearance models failed to track the target. The error of the best one tracker, MIL, was approximately three times larger than that of JISAT.

V. Quantitative Analysis

The average center position errors of the trackers for the seven test sequences are summarized in Table I. The errors colored red and blue represent the smallest and second smallest errors for each test sequence, respectively. Four different state constraints were used for JISAT: fixed-size (f), scale (sc), size (sz) and angle (a). When the tracker samples candidates, the position, and one or two constraints varies. One of the JISAT results ranks first or second place among the trackers, except for the *faceocc* sequence. However, the performance of JISAT with state constraints heavily relies on the characteristics of the sequences. LGT did not work on the *basketball* sequence. VTD tracked the *basketball* sequence for a longer period so it is marked. JISAT is similar to SPT because of the use of the superpixel as a basic component of the appearance model and of the color histogram as a feature. However, JISAT outperforms SPT in all sequences, thus indicating that the fusion of multiple features and models can effectively increase the accuracy of the estimation.

VI. Conclusions

In this paper, an efficient tracker using multiple non-rigid appearance models was proposed for tracking deformable objects. The design of the appearance model was conducted by drawing particles from the density of the appearance model in a particle filter framework. The tracking procedure was alleviated by the sampling state candidates and particles in the PMMH framework. In the experiments, the proposed method, JISAT, performed well in real-world scenarios because of its adaptability to local appearance changes and its robustness to segmentation failure.

TABLE I
Average center position errors of the trackers: fixed-size (f), scale (sc), size (sz), and angle (a) are the state constraints

seq.	[11]	[4]	[6]	[2]	[1]	[7]	[3]	[12]	[5]	JISAT			
										F	SC	SZ+A	SC+A
bask.	158	183	N/A	270	90	123	161*	279	43	29	34	37	37
div.	74	76	15	68	76	35	85	96	110	37	32	145	75
fx	69	67	70	37	43	30	31	145	N/A	18	24	45	54
gym.	N/A	108	99	62	42	29	22	41	186	25	82	16	11
face.	29	35	19	16	32	27	8	16	161	N/A	75	43	32
twin.	34	31	22	20	10	29	7	11	16	26	7	7	6
lazy.	115	165	70	57	87	72	22	83	46	N/A	42	51	33

VII. Acknowledgement

This work has been partly supported by DGIST (Daegu Gyeongbuk Institute of Science and Technology) through the project of 'Research on Improvement of Appearance Model for Visual Tracking'.

References

[1] B. Babenko, M.-H. Yang, and S. Belongie, "Visual Tracking with Online Multiple Instance Learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[2] J. Lim, D. Ross, R. sung Lin, and M. hsuan Yang, "Incremental learning for visual tracking," in *In Advances in Neural Information Processing Systems*. MIT Press, 2004, pp. 793–800.

[3] J. Kwon and K. M. Lee, "Visual tracking decomposition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 1269–1276.

[4] M. Godec, P. M. Roth, and H. Bischof, "Hough-based tracking of non-rigid objects," in *IEEE International Conference on Computer Vision*, Barcelona, Spain, 2011.

[5] S. Wang, H. Lu, F. Yang, and M.-H. Yang, "Superpixel tracking," in *IEEE International Conference on Computer Vision*, Barcelona, Spain, 2011, pp. 1323–1330.

[6] L. Cehovin, M. Kristan, and A. Leonardis, "An adaptive coupled-layer visual model for robust visual tracking," in *IEEE International Conference on Computer Vision*, Barcelona, Spain, 2011.

[7] J. Kwon and K. M. Lee, "Tracking of a non-rigid object via patch-based dynamic appearance modeling and adaptive basin hopping monte carlo sampling," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1208–1215.

[8] X. Ren and J. Malik, "Tracking as repeated figure/ground segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[9] C. Andrieu, A. Doucet, and R. Holenstein, "Particle markov chain monte carlo methods," *Journal of the Royal Statistical Society: Series B*, vol. 72, no. 3, pp. 269–342, 2010.

[10] R. Achanta, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "Slic superpixels," Ecole Polytechnique Federale de Lausanne, Tech. Rep., 2010.

[11] S. M. S. Nejhum, J. Ho, and M.-H. Yang, "Visual tracking with histograms and articulating blocks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[12] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja, "Robust visual tracking via multi-task sparse learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2042–2049.

[13] C. Y. Ren and I. Reid, "gslic: a real-time implementation of slic super-pixel segmentation," University of Oxford, Department of Engineering Science, Tech. Rep., 2011.

[14] annonymous, "untitled," in *unpublished*, 2013.

[15] Alahi, Alexandre, Ortiz, Raphael, Vandergheynst, and Pierre, "Freak: Fast retina keypoint," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 510–517.