Visual tracking using interactive factorial hidden Markov models

Jin Wook Paeng | Junseok Kwon

School of Computer Science and Engineering, Chung-Ang University, Seoul, Korea

Correspondence
Junseok Kwon, School of Computer Science and Engineering, Chung-Ang University, Seoul, Korea. Email: jskwon@cau.ac.kr

Funding information
Chung-Ang University Graduate Research Scholarship Grants, Grant/Award Number: 2019; National Research Foundation of Korea, Grant/Award Number: NRF-2020R1C1C1004907

Abstract
The authors present a novel tracking algorithm based on a factorial hidden Markov model (FHMM) that can utilise the structured information of a target. An FHMM consists of multiple hidden Markov models (HMMs), wherein each HMM aims to represent a different part of the target. Then, the geometric relation between patches is encoded in the FHMM framework via either interactive sampling or importance sampling over sets. Experimental results demonstrate that the proposed method qualitatively and quantitatively outperforms other methods, especially when the targets are highly deformable.

1 | INTRODUCTION

Visual tracking is widely used for various computer vision systems such as autonomous driving, surveillance systems, and robotics systems. Visual tracking methods can be categorised into two groups, which are generative and discriminative model-based visual trackers. Generative model-based visual trackers estimate posterior probabilities of target objects given observations from an initial frame to a current frame, in which the target objects are typically described by the models based on sparse representation [1, 2]. Then, they find a target region in the upcoming frames using the posterior probabilities. Discriminative model-based visual trackers distinguish between foreground regions from background regions by training discriminative classifiers. Then, these trained classifiers find object regions in the upcoming frames via binary classification [3–6]. In particular, correlation filters in the Fourier frequency domain have shown great success in visual tracking, because of their computational efficiency [7].

Recently, owing to the high representation ability of deep features [8], deep-learning-based visual trackers [9] have demonstrated considerably improved visual tracking accuracy and shown promising results in real-world environments. However, it remains difficult for them to track highly non-rigid targets, where the geometric structures of the targets can significantly vary over time.

This difficulty fundamentally originates from an inaccurate representation of the targets. For example, a target is typically described by a single bounding box, which is represented by a centre position \((x, y)\) and a scale \(s\). Then, the inference process is usually formulated using a single hidden Markov model (HMM) with state \(X_t = (x, y, s)\) and observation \(O_t\) for \(t = 1, \ldots, T\), as shown in Figure 1(a). However, this bounding-box representation has three degrees of freedom. Thus, it cannot take into account notable appearance changes caused by arbitrary deformation.

To handle highly deformable targets, the authors propose to represent the target as a combination of multiple patches with \(3 \times K\) degrees of freedom, wherein the state of each patch is inferred by a different HMM, as shown in Figure 1(b). To encode geometric relations among patches, all HMMs interact with each other using the proposed interaction strategies.

The contribution of this method is twofold:

- The authors propose a novel visual tracking system based on an FHMM that can efficiently infer the state of each part of the target. The proposed FHMMs can accurately describe deformable objects using multiple HMMs, in which each HMM can deal with each part of the deformable objects. In contrast, a conventional HMM fails to accurately capture deformation of the objects, because a single HMM can only describe an object as a whole and cannot exploit geometric relations between object parts.
- Two interaction strategies among multiple HMMs are presented, namely inference methods via interactive sampling
2 | RELATED WORK

In this section, visual tracking methods are introduced, which are closely related to the proposed method (i.e. visual tracking methods based on a HMM in Section 2.1, visual trackers for deformable objects in Section 2.2, visual tracking methods using deep features in Section 2.3, and generic visual tracking methods in Section 2.4).

2.1 | HMM-based visual trackers

Wang et al. [12] proposed an FHMM to fuse the results of multiple visual trackers. Yuan et al. [13] used a standard HMM to aggregate multiple visual trackers. In contrast to these methods, the proposed method uses multiple HMMs to track each part of the target. Park et al. [14] introduced an autoregressive HMM to encode the relations among varying appearances of the target over time. Ha and Kwon [15] proposed a hidden semi-Markov model (HSMM)-based visual tracker to enhance the motion model via trajectory simulation. However, these methods use a single HMM for visual tracking. In contrast to these methods, the proposed method employs multiple HMMs and has an additional mechanism that enables all HMMs to interact with each other.

Mei and Porikli [16] also presented a visual tracking method based on FHMMs. However, the FHMMs consists of only two HMMs for tracking and registration, respectively. In addition, they mutually interacted with each other. Thus, it is limited to propagate a good state of one HMM to others. In contrast, the authors’ FHMMs consists of five HMMs, which are fully interacted using explicit interaction techniques, IS and ISS. In addition, each HMM was modelled to describe each part of an object.

2.2 | Visual trackers for deformable objects

Sun et al. [17] presented a novel kernelised correlation filter that can preserve shape information for deformable objects. Lukezic et al. [18] proposed a correlation filter-based visual tracker that adopts the constellation model to describe each part of a target. Zhao et al. [19] adopted a mixture of a Gaussian model to handle partially occluded objects and introduced a deformable part model. Du et al. [20] considered dependences among parts of objects via a graph and accurately tracked deformable objects. Gao et al. [21] predicted a target location using part locations using a deep regression model. Kwon and Lee [22] proposed a Bayesian visual tracking method for highly non-rigid objects, which describes a target as multiple parts and exploits the relations between two parts. Zhaowei et al. [23] constructed a dynamic graph and matched the estimated graph with the target graph. However, these methods have no explicit interaction mechanism.

In contrast to these methods, the proposed method focuses more on the interaction among the parts of objects than

![Figure 1](image-url)  
**Figure 1** Conceptual difference between a conventional HMM and the proposed FHMM in the inference of target configurations for visual tracking: (a) conventional methods infer a target configuration $(x, y, s)$ using a single HMM; (b) the proposed method infers $K$ target configurations, $(x^k, y^k, s^k)$ for $k = 1, \ldots, K$, using the proposed FHMM, which consists of $K$ HMMs. In the proposed method, each HMM aims to track a part of the target, while all HMMs communicate with each other via interaction processes. FHMM, factorial hidden Markov model; HMM, hidden Markov model.
conventional methods and formulates two interacting algorithms in the FHMM framework.

2.3 Visual trackers using deep features

Deep neural networks [24–26] enabled conventional visual tracking methods to extract more representative features, which significantly improves the accuracy of visual trackers, as demonstrated by [27–31]. Ma et al. [27] presented hierarchical convolutional features to represent a target as multiple levels of abstraction and accurately track the target. Nam et al. [28] transformed visual tracking problems into binary classification ones and improved the visual tracking accuracy using multi-domain features. Wang et al. [29] introduced a sequential training strategy for online visual tracking. Wang and Yeung [30] presented a stacked denoising autoencoder to extract deep features and classified each image patch into foreground or background. Kwon et al. [31] integrated deep neural networks into Monte Carlo sampling to infer multiple variables simultaneously. However, these methods did not explicitly deal with highly deformable objects.

By contrast, the proposed method adopted deep features for a novel FHMM to deal with deformable objects.

2.4 Generic visual tracking methods

Isard and Blake [32] introduced a Bayesian framework based on particle filtering for visual tracking, which approximated non-Gaussian distributions using particles. Khan et al. [33] presented a Markov chain Monte Carlo-based visual tracking, where high-dimensional posterior distributions were efficiently estimated. Kwon et al. [34] proposed an uncertainty calibrated MCMC, in which the uncertainty in the Monte Carlo sampling was minimised for accurate inference of posterior distributions. Gustafsson et al. [35] presented a deep probabilistic regression method based on Monte Carlo sampling to approximate the likelihood distribution and optimise an energy model for visual tracking.

In contrast to these methods, the proposed method accurately estimated a target posterior distribution using multiple Markov chains and can prevent the visual tracker from becoming trapped in local optima using multiple Markov chains.

3 VISUAL TRACKING BY FHMM

In HMM, a target state $X$ is represented as a set of states whose length is $T$, that is $X = \{X_1, \ldots, X_T\}$, while an observation $O$ is expressed as a set of image features within bounding boxes across $T$ frames, that is $O = \{O_1, \ldots, O_T\}$. Thus, the number of states is equivalent to the number of observations. Then, the joint distribution for the HMM is formulated as

$$p(O, X) = \prod_{t=1}^{T} p(O_t | X_t) \prod_{t=1}^{T} p(X_t | X_{t-1}),$$

where $p(O_t | X_t)$ and $p(X_t | X_{t-1})$ denote the likelihood and transition probability of the HMM at time $t$, respectively.

### Algorithm 1 FHMM-IS

| Input: $\hat{X}^{(k)}$ for $k = 1, \ldots, K$ at the previous frame. |
|---------------------------------------------------------------|
| Output: $\hat{X}^{(k)}$ for $k = 1, \ldots, K$ at the current frame. |
| 1: for $k = 1$ to $K$ (for each part) do |
| 2: for $t = 1$ to $T$ do |
| 3: Move to the next state using $p(X_t^{(k)} | X_{t-1}^{(k)})$ in (4). |
| 4: Compute the likelihood using $p(O_t | X_t^{(k)})$ in (3). |
| 5: Sample $\theta \sim U[0, 1]$. |
| 6: if $\theta < \gamma$ then |
| 7: Propagate $X_t^{(k)}$ with probability |
| 8: end if |
| 9: $\gamma = 1 - \frac{0.5}{T} \times t$. |
| 10: end for |
| 11: end for |
| 12: Find the best configuration $\hat{X}^{(k)}$ using (5). |

An FHMM [36] adopts $K$ standard HMMs. In an FHMM, $X_t^{(k)}$ denotes the $t$-th state of the $k$-th HMM. Each HMM is expressed as a set of states whose length is $T$, that is $X^{(k)} = \{X_1^{(k)}, \ldots, X_T^{(k)}\}$. Given the observations $O^{(k)} = \{O_1^{(k)}, \ldots, O_T^{(k)}\}$, the joint distribution for the $k$-th HMM is formulated as

$$p(O^{(k)}, X^{(k)}) = \prod_{t=1}^{T} p(O_t^{(k)} | X_t^{(k)}) \prod_{t=1}^{T} p(X_t^{(k)} | X_{t-1}^{(k)}),$$

where $p(O_t^{(k)} | X_t^{(k)})$ and $p(X_t^{(k)} | X_{t-1}^{(k)})$ denote the likelihood and transition probability of the $k$-th HMM at time $t$, respectively. The likelihood indicates the similarity between the observed feature $O_t^{(k)}(X_t^{(k)})$ (i.e. features in the region described by $X_t^{(k)}$) and the reference feature $O_{ref}^{(k)}$ in terms of the distance metric $D$:

$$p(O_t^{(k)} | X_t^{(k)}) = e^{-D(O_t^{(k)}, O_{ref}^{(k)})},$$

where deep features are extracted from the 14th feature map of the VGG-19 network [8] as the observed features, which was pre-trained by the ImageNet dataset [37]. We adopt the correlation filter used in [9] as the distance metric. Then, we design the transition probability as follows:

$$p(X_t^{(k)} | X_{t-1}^{(k)}) = \mathcal{N}(X_t^{(k)} | X_{t-1}^{(k)}, \sigma^2),$$
where a new state $X_t^{(k)}$ is proposed by the normal function with mean $X_{t-1}^{(k)}$ and variance $\sigma^2$. In all experiments, $\Sigma$ is fixed to the diagonal matrix with diagonal entries $(\sigma_x, \sigma_y, \sigma_z)$, where $\sigma_x$, $\sigma_y$, and $\sigma_z$ denote the variances for $x$, $y$, and $z$ scale coordinates, respectively. Note that $\sigma_x$, $\sigma_y$, and $\sigma_z$ hyperparameters, which are set to 0.01, 0.01, and 0.1, respectively.

The goal of an FHMM is to find the best state $\hat{X}^{(k)}$ for each HMM (i.e., for each part of the target) that maximises the likelihood $p\left(O_t^{(k)}\mid X_t^{(k)}\right)$ in (3):

$$\hat{X}_t^{(k)} = \arg\max_{X_t^{(k)}} p\left(O_t^{(k)}\mid X_t^{(k)}\right), \quad (5)$$

for $k = 1, \ldots, K$ and $t = 1, \ldots, T$.

**Algorithm 2 FHMM-ISS**

**Input:** $\hat{X}_t^{(k)}$ for $k = 1, \ldots, K$ at the previous frame.  
**Output:** $\hat{X}_t^{(k)}$ for $k = 1, \ldots, K$ at the current frame.

1. for $k = 1$ to $K$ (for each part) do
2. for $t = 1$ to $T$ do
3. Move to the next state using $p\left(X_t^{(k)}\mid X_{t-1}^{(k)}\right)$ in (4).
4. Compute the likelihood using $p\left(O_t^{(k)}\mid X_t^{(k)}\right)$ in (3).
5. Estimate $Z_t$ using $E\{\sigma(z)\}$ [inverse of $Z_t$] in (7).
6. Normalise $p\left(O_t^{(k)}\mid X_t^{(k)}\right)$ by $\frac{p\left(O_t^{(k)}\mid X_t^{(k)}\right)}{\sum_{k=1}^{K} p\left(O_t^{(k)}\mid X_t^{(k)}\right)}$ for $k = 1, \ldots, K$.
7. end for
8. end for
9. Find the best configuration $\hat{X}^{(k)}$ using (5).

4 | **INFERECE VIA IS**

The proposed IS method is an extension of the interactive Markov chain Monte Carlo method [10], where multiple HMMs can communicate with each other about the accurate state of each part. Using IS, one HMM can communicate with other HMMs regarding the good states of target objects. Because each HMM infers different states, this communication process can induce the fusion of all states with different weights. A state has a large weight and the corresponding HMM frequently propagates the states to other HMMs, if the state makes a large likelihood value. Then, a HMM interacts with other HMMs and makes leaps to better states of target objects. The IS method has two modes. In the normal mode, each HMM proceeds separately. However, in the interaction mode, the $k$-th HMM propagates its state to the other HMMs, which substitute their states by the state of the $k$-th HMM with probability $a_{tk}^j$:

$$a_{tk}^j = \frac{p\left(O_t^{(k)}\mid X_t^{(k)}\right)}{\sum_{k=1}^{K} p\left(O_t^{(k)}\mid X_t^{(k)}\right)}$$

where $p\left(O_t^{(k)}\mid X_t^{(k)}\right)$ is the likelihood of the $k$-th HMM at time $t$. Note that the state of the $k$-th HMM cannot accurately describe the states of other parts, although all parts of a single target are located close to each other. However, the state of the $k$-th HMM prevents the states of other parts from drifting into the background. The proposed IS method executes the interaction mode with probability $\gamma = 1 - a_{tk}^j$, where $t$ and $T$ denote the iteration index and the total number of iterations, respectively. The value of $\gamma$ linearly decreases, as the sampling proceeds. Figure 2 conceptually illustrates the IS method for the interaction among multiple HMMs. Algorithm 1 shows the whole process of the proposed visual tracking method with IS (i.e., FHMM-IS).

5 | **INERENCE VIA ISS**

The proposed ISS for FHMM is a variant of the one reported in [11]. Using ISS, each FHMM calibrates the corresponding likelihood value using the partition function $Z_t$.

With this modification, each HMM is affected by other HMMs. For this, ISS defines a set of states and estimates a statistic of the set (i.e., $Z_t$), which encodes the likelihoods of all HMMs at time $t$.
Then, using $Z_t$, the likelihood of each HMM at time $t$ is modified. If the contribution of $p(O_t^{(k)}|X_t^{(k)})$ to $Z_t$ is minor, $p(O_t^{(k)}|X_t^{(k)})$ is underweighted by $\frac{1}{Z_t}$. If the contribution of $p(O_t^{(k)}|X_t^{(k)})$ to $Z_t$ is large, $p(O_t^{(k)}|X_t^{(k)})$ is overweighted by $\frac{1}{Z_t}$. Thus, a HMM can interact with other HMMs and make leaps to better states of target objects. To compute $Z_t$ in (7), a set-proposal function $q$ is designed that proposes a set of new states based on a set of previous states. Then, the transition probability in (4) can be substituted by $q$ as follows:

$$q \left( \left\{ X_t^{(k)} \right\}, \left\{ X_{t-1}^{(k)} \right\} \right) = g \left( \left\{ X_t^{(k)} \right\}, \left\{ X_{t-1}^{(k)} \right\}, \sigma^2 \right) = \prod_{k=1}^{K} N \left( X_t^{(k)}; X_{t-1}^{(k)}, \sigma^2 \right),$$

where $g$ is a multivariate Gaussian function and $N$ is defined in (4). The probability of $X_t$ being included in the set $\left\{ X_t^{(k)} \right\}$ can be computed, and is sampled by $q$, as follows:

$$\beta(X_t) = \sum I \left( X_t \in \left\{ X_t^{(k)} \right\} \right) q \left( \left\{ X_t^{(k)} \right\} \right),$$

where the function $I$ outputs unity, if $X_t \in \left\{ X_t^{(k)} \right\}$. Using $\beta(X_t)$, the unbiased partition function $Z_t$ in (7) can be estimated. The entire proof can be found in [11]. Figure 3 conceptually illustrates the ISS method for the interaction among multiple HMMs. Algorithm 2 shows the whole process of the proposed visual tracking method with IS (i.e. FHMM-ISS).

### 6 | EXPERIMENTS

The authors used five parts of the target (i.e. $K = 5$) and 100 samples for each part (i.e. $T = 100$), and randomly initialised the configuration of each part. The proposed methods were compared to state-of-the-art deep learning-based methods, namely SiamRPN++ [38], TADT [39], DAT [40], SiamDW [41], SINT [42], SINT-op [42], C-COT [43], ECO [44], and ECO-HC [44]. These trackers were evaluated through 50 test sequences in the OTB dataset [45]. The OTB test data set is a standard visual tracking benchmark data set and contains either 50 (OTB-50) or 100 (OTB-100) challenging sequences with several different attributes, which are out-of-view objects, low resolution, background clutter, out-of-plane rotation, motion blur, fast motion, deformations, occlusions, in-plane rotation, scale variation, and varying illumination. Visual trackers were evaluated using the precision, success plots, and AUC [45]. The precision plot is the percentage of frames, where the distance between the bounding box estimations and ground truth is less than a threshold. The success plot is the percentage of frames, where intersection of union between the bounding box estimations and ground truth is greater than a threshold. The AUC is the area under the success plot. The authors conducted all experiments using Intel CPU i7 3.60 GHz and GeForce RTX 2080.

#### 6.1 | Ablation study

Figure 4 qualitatively compares the proposed methods using the OTB-50 dataset. Blue, red, yellow, green, and white boxes show the states of five parts of the target. As shown in the figure, FHMM-ISS accurately localised all parts via the interaction process, that is ISS.

The method was analysed in a component-wise manner using the OTB-50 data set and verified the sensitiveness of several hyperparameters used in the proposed method. Thus, the proposed method was not compared with other methods here, but three variants of the proposed method were developed, namely FHMM, FHMM-IS, and FHMM-ISS. FHMM denotes the baseline tracker, which employs the proposed factorial HMM as an inference method. FHMM-IS and FHMM-ISS denote improved versions of the baseline tracker, which is combined with interactive sampling and importance sampling over sets, respectively. As shown in Figure 5, FHMM-IS and FHMM-ISS outperforms FHMM, thereby demonstrating the effectiveness of the proposed interaction strategies for multiple HMMs. The interaction method based on ISS was the best in terms of precision and success plots.
Figure 6 demonstrates the effectiveness of the proposed interaction strategy, ISS. At the frame #99, the white-coloured FHMM failed to track a certain part of the target and became drifted into the background. However, the proposed method prevented the white-coloured FHMM from being further drifted and recovered its state with the help of other FHMMs using the proposed interaction mechanism, ISS.

Table 1 shows the tracking accuracy of the proposed method in terms of AUC, as the number of parts (i.e. FHMMs) increases. If a single part or FHMM is used, the tracking accuracy dropped considerably, which demonstrates the effectiveness of using multiple FHMMs. However, using many parts or FHMMs also deteriorated the visual tracking performance in terms of speed and accuracy. Nevertheless, the proposed FHMM-ISS was not significantly sensitive to the number of parts. Table 2 shows the tracking accuracy of the proposed method in terms of AUC, as the number of samples (i.e. the number of states in FHMMs) increases. As shown in Table 2,
TABLE 1 Visual tracking performance of the proposed FHMM-ISS according to the number of parts (i.e. $K$) in terms of AUC

| $K$ | 1 | 3 | 5 | 10 | 15 |
|-----|---|---|---|----|----|
| AUC | 0.610 | 0.685 | 0.701 | 0.698 | 0.675 |

Abbreviations: FHMM, factorial hidden Markov model; ISS, importance sampling over sets.

TABLE 2 Visual tracking performance of the proposed FHMM-ISS according to the number of samples (i.e. $T$) in terms of AUC

| $T$ | 10 | 50 | 100 | 200 | 300 |
|-----|----|----|-----|-----|-----|
| AUC | 0.678 | 0.695 | 0.701 | 0.701 | 0.703 |

Abbreviations: FHMM, factorial hidden Markov model; ISS, importance sampling over sets.

TABLE 3 Visual tracking performance of the proposed FHMM-ISS according to $K \times T$ in terms of speed (i.e. FPS)

| $K \times T$ | 300 | 400 | 500 | 600 | 700 |
|-------------|-----|-----|-----|-----|-----|
| AUC | 93 | 86 | 79 | 71 | 65 |

Abbreviations: FPS, frames per seconds.

TABLE 4 Computational costs of FHMM in terms of FPS

| SiamRPN++ | ECO | COT | SINT | FHMM |
|-----------|-----|-----|------|------|
| FPS | 35 | 7 | 1 | 5 | 79 |

Abbreviations: FHMM, factorial hidden Markov model; FPS, frames per second.

similar results were obtained even with a different number of samples, which implies that the effectiveness of the proposed method mainly stems from multiple FHMMs and novel interacting mechanisms. Table 3 shows the speed of this method in terms of frames per second. The proposed FHMM-ISS performed in real time, although the authors used 700 samples in total. The proposed interaction strategies and inference processes of FHMMs do not require high computation costs. Table 4 evaluates the computational costs of recent state-of-the-art visual trackers using the OTB data set. The visual tracker runs in real time and is the fastest algorithm, because the proposed method does not adopt complex deep neural network architectures.

6.2 | Quantitative comparison

The authors quantitatively compared the proposed method with non-deep-learning-based trackers using the OTB-50 data set, as shown in Figure 7. This method considerably outperformed conventional visual trackers. In particular, the proposed FHMMs can accurately describe deformable objects using multiple HMMs, in which each HMM can deal with each part of the deformable objects. Note that a conventional HMM cannot accurately capture deformation of the objects, because it cannot exploit geometric relations between object parts.

The authors quantitatively compared the proposed method with recent deep-learning-based trackers using the OTB-50 data set, as shown in Figure 8. Although the visual trackers do not employ deep neural architectures and only utilise deep features, the proposed FHMM-ISS is comparable with state-of-the-art visual trackers in terms of both precision and success rate. In particular, the proposed visual tracker accurately tracked highly deformable objects by modelling multiple parts using FHMMs and interacting FHMMs. With high probability, at least one FHMM can accurately track the corresponding part of a target and its state can be used to re-initialise the states of other FHMMs. SiamDW and ECO were the best in terms of precision and success plots, respectively. However, these trackers adopted deep neural network architectures, which required high computation costs.

Table 5 quantitatively compared the proposed method with other trackers using the OTB-100 data set. The proposed FHMM-ISS show the second-best performance in terms of AUC. ECO shows state-of-the-art performance, but cannot run in real time. In contrast, the proposed FHMM-ISS is considerably faster than ECO, as shown in Table 4. In addition, the authors compared this method with JTR [16]. As shown in Table 5, the proposed method considerably outperforms JTR, because the proposed method fully communicates with multiple HMMs and each HMM handles changes in each part of an object.

6.3 | Qualitative comparison

Figures 9–12 show qualitative results of visual tracking methods. In most test videos, severe deformation of the target exists. In addition, the MotorRolling sequence contains objects which are severely rotated over time, while the Ironman sequence includes deformable objects with considerable changes in illumination conditions. The Diving sequence has highly non-rigid objects, where the object size and aspect ratio considerably vary over time. The Shaking sequence contains faces, which significantly shake. The Soccer and Matrix sequences have diverse obstacles at the same time, which include illumination variation, scale variation, occlusion, motion blur, and background clutters. The Skating1 sequence contains non-rigid object deformation and considerable illumination variation. The Bird1 sequence has fast motion, deformation, and out-of-view attributes. As shown in Figure 8, the proposed FHMM-ISS accurately tracked highly deformable objects. Each FHMM robustly tracked each part of the targets, while the proposed interaction strategies prevent each part from being drifted into the background. Moreover, each FHMM can handle different obstacles in the scene and propagate information on good states to other FHMMs. SiamSW and ECO also tracked the targets successfully. However, they failed to represent the target size and aspect ratio precisely.

7 | CONCLUSION

In this study, a novel visual tracking method was developed based on an interactive FHMM that can utilise the structured
**FIGURE 7** Quantitative comparison with non-deep-learning-based methods

**FIGURE 8** Quantitative comparison with deep-learning-based methods. FHMM, factorial hidden Markov model

**TABLE 5** Quantitative comparison using the OTB-100 data set in terms of AUC

| SCM | Struck | CXT | VTS | VTD | LSK | OAB | MIL | CSK | ECO | DAT | SiamRPN+ | TADT | COT | SiamDW | SINT | JTR | FHMM-ISO |
|-----|--------|-----|-----|-----|-----|-----|-----|-----|-----|-----|---------|------|-----|--------|------|-----|---------|
| AUC | 0.447  | 0.461| 0.412| 0.363| 368 | 0.382| 0.364| 0.219| 0.381| 0.694| 0.668   | 0.637 | 0.660| 0.671  | 0.670 | 0.574| 0.368   | 0.689 |

**FIGURE 9** Qualitative comparison with deep-learning-based methods

MotorRolling  Ironman  Diving  Shaking
ACKNOWLEDGEMENTS

This work was partly supported by the Chung-Ang University Graduate Research Scholarship Grants in 2019 and also partly supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (NRF-2020R1C1C1004907).

REFERENCES

1. Ross, D.A., et al.: Incremental learning for robust visual tracking. Int. J. Comput. Vis. 77(1–3), 125–141 (2008)
2. Mei, X., Ling, H.: Robust visual tracking using l1 minimisation. ICCV (2009)
3. Henriques, J.F., et al.: High-speed tracking with kernelised correlation filters. IEEE Trans. Pattern Anal. Mach. Intell. 37(3), 583–596 (2015), https://doi.org/10.1109/tpami.2014.2345390
4. Danelljan, M., et al.: Learning spatially regularised correlation filters for visual tracking. CVPR (2015)
5. Zhang, J., Ma, S., Sclaroff, S.: MEEM: robust tracking via multiple experts using entropy minimisation. ECCV (2014)
6. Kalal, Z., Mikolajczyk, K., Matas, J.: Tracking-learning-detection. IEEE Trans. Pattern Anal. Mach. Intell. 34(7), 1409–1422 (2012)
7. Bolme, D.S., et al.: Visual object tracking using adaptive correlation filters. CVPR (2010)
8. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. ICLR (2015)
9. Tao, R., Gavves, E., Smeulders, A.W.M.: Siamese instance search for tracking. CVPR (2016)
10. Corander, J., Ekdahl, M., Koski, T.: Parallel interacting mcmc for learning of topologies of graphical models. Data Min. Knowl. Disc. 17(3), 431–456 (2008)
11. Hadjis, S., Ermon, S.: Importance sampling over sets: a new probabilistic inference scheme. UAI (2015)
12. Wang, N., Yan Yeung, D.: Ensemble-based tracking: Aggregating crowdsourced structured time series data. ICML (2014)
13. Yuan, Y., et al.: Visual object tracking by structure complexity coefficients. IEEE Trans. Multimedia. 17(8), 1125–1136 (2015)
14. Park, D.W., Kwon, J., Lee, K.M.: Robust visual tracking using autoregressive hidden markov model. CVPR (2012)
15. Ha, S., Kwon, J.: Visual tracking enhancement by trajectory simulation based on hidden semi-Markov model. Electron Lett. 56(2), 85–87 (2020)
16. Mei, X., Porikli, F.: Joint tracking and video registration by factorial hidden markov models. ICASSP (2008)
17. Sun, X., et al.: Non-rigid object tracking via deformable patches using shape-preserved kcf and level sets. ICCV (2017)
18. Lukeziec, A., Zach, I.C., Kristan, M.: Deformable parts correlation filters for robust visual tracking. IEEE Cybernetics. 48(6), 1849–1861 (2017)
19. Zhao, C., et al.: Learning weighted part models for object tracking. Comput. Vis. Image Understand. 143(2), 173–182 (2016)
20. Du, D., et al.: Online deformable object tracking based on structure-aware hyper-graph. IEEE Trans. Image Process. 25(8), 3572–3584 (2016)
21. Gao, J., et al.: P2T: Part-to-target tracking via deep regression learning. IEEE Trans. Image Process. 27(6), 3074–3086 (2018)
22. Kwon, J., Lee, K.M.: Highly nonrigid object tracking via patch-based dynamic appearance modelling. IEEE TPAMI. 35(10), 2427–2441 (2013)
23. Zhao, C., et al.: Robust deformable and occluded object tracking with dynamic graph. IEEE TIP. 23(12), 5497–5509 (2014)
24. LeCun, Y., et al.: Gradient-based learning applied to document recognition. Proc. IEEE. 86(11), 2278–2324 (1998)
25. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. NIPS (News Physiol Sci) (2012)
26. Ren, S., et al.: Faster r-cnn: towards real-time object detection with region proposal networks. NIPS. (2015)
27. Ma, C., et al.: Hierarchical convolutional features for visual tracking. ICCV (2015)
28. Nam, H., Han, B.: Learning multi-domain convolutional neural networks for visual tracking. CVPR (2015)
29. Wang, L., et al.: Sct: sequentially training convolutional networks for visual tracking. CVPR (2016)
30. Wang, N., Yeung, D.Y.: Learning a deep compact image representation for visual tracking. News Physiol. Sci. (2013)
31. Kwon, J.: Robust visual tracking based on variational auto-encoding Markov chain Monte Carlo. Inf. Sci. 512, 1308–1323 (2020)
32. Isard, M., Blake, A.: Icondensation: Unifying low-level and high-level tracking in a stochastic framework. ECCV (1998)
33. Khan, Z., Balch, T., Dellaert, F.: MCMC-based particle filtering for tracking a variable number of interacting targets. IEEE TPAMI. 27(11), 1805–1918 (2005)
34. Kwon, J.: Uncertainty calibrated Markov chain Monte Carlo sampler for visual tracking based on multi-shape posterior. J. Math. Imaging Vis. 60(5), 681–691 (2018)
35. Gustafsson, F.K., et al.: Energy-based models for deep probabilistic regression. ECCV (2020)
36. Martens, K., Titsias, M., Yau, C.: Augmented ensemble meme sampling in factorial hidden markov models. PMLR. 89, 2359–2367 (2019)
37. Deng, J., et al.: ImageNet: a large-scale hierarchical image Database. CVPR (2009)
38. Li, B., et al.: SiamRPN++: Evolution of siamese visual tracking with very deep networks. CVPR (2018)
39. Li, X., et al.: Target-aware deep tracking. CVPR (2019)
40. Pu, S., et al.: Deep attentive tracking via reciprocal learning. NIPS. (2018)
41. Zhang, Z., Peng, H.: Deeper and wider siamese networks for real-time visual tracking. CVPR (2019)
42. Tao, R., Gavves, E., Smeulders, A.W.M.: Siamese instance search for tracking. CVPR, (2016)
43. Danelljan, M., et al.: Beyond correlation filters: learning continuous convolution operators for visual tracking. ECCV (2016)
44. Danelljan, M., et al.: ECO: Efficient convolution operators for tracking. CVPR (2017)
45. Wu, Y., Lim, J., Yang, M.H.: Online object tracking: a benchmark. CVPR (2013)

How to cite this article: Paeng JW, Kwon J. Visual tracking using interactive factorial hidden Markov models. IET Signal Process. 2021;1–10. https://doi.org/10.1049/si2.12037