A NOVEL ADAPTIVE KERNEL CORRELATION FILTER TRACKER WITH MULTIPLE FEATURE INTEGRATION

Zhonggeng Liu, Zhichao Lian, Yang Li

School of Computer Science and Engineering, Nanjing University of Science and Technology, China
15951939370@163.com, lzcts@163.com,694437416@qq.com

ABSTRACT

Recently, correlation filters (CFs) for visual tracking present competitive performances on both accuracy and robustness, but there is still a need for improving their overall tracking capabilities. Most CF trackers learn a best filter to regress training data to a fixed target response, which might lead to drifting. In this paper, we present an appealing tracker based on the Kernelized Correlation Filter (KCF), which can adaptively change the target response. Furthermore, we utilize a fast and accurate scale estimation approach by learning an independent correlation filter instead of employing an exhaustive scale search strategy to estimate the target size. In addition, color naming integrates into the histogram of orientation gradient feature to further boost the performance for our tracker. We validate our tracker on the popular OTB50 datasets, which outperforms the state-of-the-art methods in terms of efficiency and accuracy.

Index Terms—Visual Tracking, Correlation Filter, Adaptive Target Response, Scale Estimation, Feature integration.

1. INTRODUCTION

Visual tracking is a classical problem in computer vision with various applications, such as surveillance, robotics and automation. In general, the task of visual object tracking is to estimate the trajectory of the target throughout the video frames given an initial location of the target in the first frame. The problem is challenging due to occlusions, scale variations, fast motion, motion blur, etc.

Recently, Correlation filters (CFs) [1][2][3][4][5] have received much attention in visual tracking due to their state-of-the-art tracking performance both in speed and accuracy. CFs based methods have shown competitive performance on the OTB dataset [6] and the VOT 2014 Challenge [7]. Bolme et al. [1] and Heriques et al. [2] introduced the correlation filter framework into visual tracking. They worked by learning an optimal correlation filer used to localize the object in the next frame by identifying the location of maximal correlation response. The tracking framework is computational efficiency due to exploiting the circulant structure both in training and detection step. In [2][3], the KCF tracker was obtained by solving a Ridge Regression problem to regress the training data to a fixed target response which is a Gaussian peak at the ground truth object location in the first frame. In [8], Adel Bibi et al. found that once the detection step of the tracker is inaccurate, or circular shifts cannot represent real translations as shown in Fig.1, this error will be propagated to the newly computed filter for updated scheme. Finally, the tracker might drift.

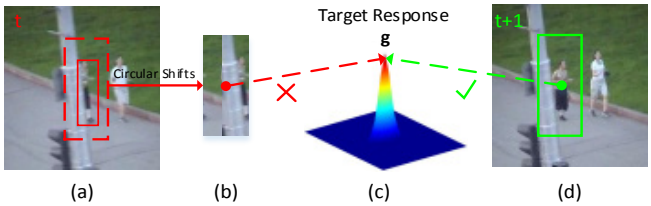


Fig.1. Circular shifts do not approximate to real translations. The red box in (a) is the search window at frame t, and patch (b) constructed by circular shifts cannot represent real translations in case of occlusions because both the occluder and target are shifted, while the blue box in (d) is real translations. This error will infect the detection and training step of CF based tracker.

In this paper, we aim to build a novel real-time target response adaptive and scale adaptive KCF tracker with multiple feature integration. Inspired by [8], we add a target response adaptive paradigm to our tracker, which is able to handle errors in the detection step and recover from drift. In addition, considering that KCF [2][3] trackers cannot handle the scale changes of targets, we train another scale correlation filter to estimate the scale change, which uses the histogram of orientation gradient (HOG) feature of the target at a set of different scales instead of exhaustive scale searching. This scale estimation method has been proven faster and more robust in [4], which is based on MOSSE tracker [1]. Finally, to boost the performance further, our tracker integrates the HOG and color naming (CN) [14] which can handle more challenge scenarios. Extensive experimental results demonstrate that our tracker outperforms the state-of-the-art trackers on benchmark datasets [6] with 51 video sequences.

Our main contributions of this work are summarized as follows: 1) We design a more realistic target response to enhance the robustness of KCF tracker. We reformulate the optimization to efficiently solve for both the best translation correlation filter and target response jointly, whereby the

target response is regularized using correlation scores evaluated at sampled translations instead of circular shifts. 2) We extend the KCF tracker to be able to deal with scales variations, which is computed fast and accurately. 3) We adopt a multiple feature integration, which combines the HOG and CN to strengthen the discriminative ability of the KCF tracker. The proposed tracker is not only real-time but also superior to the state-of-the-art trackers in both accuracy and robustness.

2. RELATED WORK

Most visual object tracking methods in general can be divided into two main categories: generative and discriminative methods. The main aim of the generative tracker is to search for the target which is the most similar in appearance to generative model, while discriminative trackers formulate the tracking task as a binary classification problem to predict target location that is the most distinctive from the background. Since our tracker belongs to discriminative trackers, we restrict our review to those trackers close to our work, such as TLD [9], Struck [10], CSK [2], MOSSE [1], KCF [3], DSST [4], SAMF [5], etc.

Discriminative approaches employ machine learning approaches to classify the target and background. TLD [9] used a boosted classifier to predict the target location and scale. Struck [10] used a kernelized structured output Support Vector Machine (SVM) to distinguish the target and background. Although these trackers aforementioned achieved good performance in the benchmark datasets [6], their computational cost is very high to limit their speed. Recently, the correlation filter-based trackers [1][2][3][4][5] have been studied due to their appealing performance both in speed and accuracy. In [1], Bolme et al. designed a correlation filter named MOSSE, which performed similarly to the most sophisticated trackers of the time but reached a runtime of 600 FPS. In [2][3], Henriques et al. utilized the circulant matrix and DFT to efficiently solve a linear regression problem in the Fourier domain. Many improvements have been made to this KCF tracker. In [11], Ma et al. trained a random fern classifier to re-detect objects in case of tracking failure, which achieved a long-term tracking. Liu et al. [12] combined part-based trackers and correlation filters together to achieve real-time robust tracking. Bibi et al. [13] developed a KCF scheme that incorporated multiple multi-dimensional templates in training. In [5], Li et al. proposed a scale adaptive scheme to deal with the fixed template size in KCF tracker, and they fused HOG and CN features to boost the performance for their tracker.

3. OUR TRACKER

In this section, we first review the KCF [3] tracker. Secondly, we introduce the target response adaptive scheme. Furthermore, we discuss our fast and accurate scale estimation method. Finally, the powerful features, fused HOG and CN, are proposed to enhance the discriminant power of our tracker.

3.1. Kernelized Correlation Filter

KCF trackers [2][3] aim to find a best filter $\boldsymbol{w}$ that minimizes the squared error over circulant samples $\boldsymbol{X}$:

$$\min_{\boldsymbol{w}} ||\boldsymbol{X}\boldsymbol{w} - \boldsymbol{g}||^2 + \lambda||\boldsymbol{w}||^2 \quad (1)$$

where λ is a regularization parameter and $\boldsymbol{g}$ is the hand-crafted target response which is a Gaussian peak at the ground truth object location in the first frame. In [3], Henriques et al. used the kernel trick [18] to allow non-linear regression, $\boldsymbol{w} = \sum_i \alpha_i \emptyset(\boldsymbol{x_i})$, where $\emptyset(.)$ denotes the mapping to a kernel space, and the coefficient $\boldsymbol{\alpha}$, defined by (2), is alternative representation in the dual space.

$$\mathcal{F}(\boldsymbol{\alpha}) = \frac{\mathcal{F}(\boldsymbol{g})}{\mathcal{F}\left(\emptyset^T(\boldsymbol{x})\emptyset(\boldsymbol{x})\right)+\lambda} \quad (2)$$

where $\mathcal{F}(.)$ denotes the DFT operator. In the detection step, given the candidate image patch $\boldsymbol{z}$, the target location is estimated by maximum value of $\widehat{\boldsymbol{f_z}}$ which is a vector, containing the output for all cyclic shifts of $\boldsymbol{z}$.

$$\widehat{\boldsymbol{f_z}} = \mathcal{F}^{-1}(\mathcal{F}(\boldsymbol{\alpha}) \odot \mathcal{F}\left(\emptyset^T(\boldsymbol{z})\emptyset(\boldsymbol{x})\right)) \quad (3)$$

3.2. Adaptive Target Responses

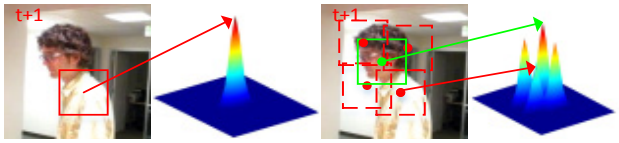
As mentioned earlier, since the detection step of CFs trackers might be error or circular shifts cannot correspond to real translations as shown in Fig.1, the tracker might drift. Therefore, we design an adaptive target response $\boldsymbol{g}$ in every frame similar to the idea in [8] instead of a fixed one constructed at the first frame. The peak values of $\boldsymbol{g}$ uses real translation measurements of correlation, which combines the target appearance information and prior motion information. Thus, we obtain the following joint optimization to solve the best filter $\boldsymbol{w}$ and target response $\boldsymbol{g}$:

$$\min_{\boldsymbol{w},\boldsymbol{g}} ||\boldsymbol{X}\boldsymbol{w} - \boldsymbol{g}||^2 + \lambda||\boldsymbol{w}||^2 + \xi||\boldsymbol{g} - \boldsymbol{g_0}||^2 \quad (4)$$

where $\boldsymbol{g}$ is assumed to follow the noise model: $\boldsymbol{g} = \boldsymbol{g_0} + \boldsymbol{\varepsilon}$, and $\boldsymbol{g} \sim N(\boldsymbol{g_0}, diag^{-1}(1/2\xi))$.

In the training step, given a new frame t+1, we firstly sample m translations at which surround the previous estimated position. Then we compute the correlation scores at these m positions, which is used to fill the corresponding m entries in $\boldsymbol{g_0}$. Finally, we employ Gaussian interpolation to compute the rest of entries in $\boldsymbol{g_0}$. The best filter $\boldsymbol{w}$ is shown as follow and * denotes the complex-conjugate:

$$\mathcal{F}(\boldsymbol{w}) = \frac{\xi(\mathcal{F}(\boldsymbol{x}) \odot \mathcal{F}(\boldsymbol{g_0}))}{\xi(\mathcal{F}(\boldsymbol{x}) \odot \mathcal{F}(\boldsymbol{x})^*)+\lambda(1+\xi)} \quad (5)$$



(a) (b)

Fig.2. Shows the difference between KCF trackers (a) and our tracker (b) in the training step.

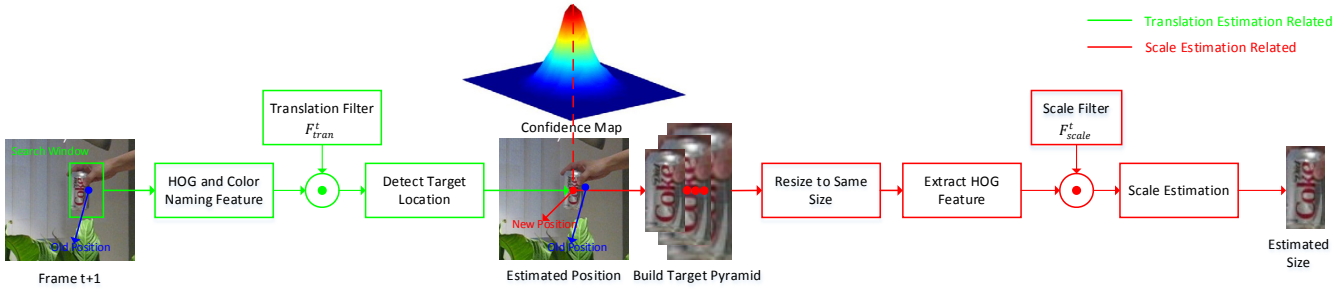


Fig.3. Flowchart of the proposed scale adaptive KCF tracker. The tracker is decomposed into translation filter and scale filter: we use the target and its surrounding patch to extract the HOG and CN feature to estimate translation, and scale is estimated by the target appearance information.

3.3 Feature Integration

Since the HOG [16] focuses on the image gradient information while color feature pays attention to color information, the two features are complementary to each other. In this work, we adopt HOG and CN [14] integration scheme to boost our tracker's performance.

Histogram of Orientation Gradient (HOG) is one of most popular features in visual tracking. HOG extracts the image gradient information and it is effective in object detection. Dalal et al. [17] designed a 36-dimensional HOG feature. In order to compute faster, we use an alternative 31-dimensional HOG described in [16], which has a more powerful discrimination.

Color names (CN) is the action of assigning linguistic color labels to colors in the world and it is widely used in computer vision such as image retrieval, object detection [20], object recognition [21] and action recognition [22]. Belin and Kay [19] obtained 11 basic linguistic color labels. We use the mapping provided by [14] to convert RGB images to 11-dementiosnal color names.

In this paper, the CN is fused into HOG result in obtaining 42-dimensional features, and note that the size of feature map is fixed. We evaluated the discriminant power of these features in OTB50 datasets, which proved HOG and CN integration had the most appealing performance in visual tracking.

3.4 Scale Adaptive

KCF tracker can only detect the object position and cannot deal with the object scale changes as described in Section 3.1. A common approach for detecting an object at different scales is to apply a classifier at multiple scales [5][12][13], which increases the computational cost of the tracker. In order to compute faster and track more accurately, we learn two correlation filters, one for translation estimation and another for scale estimation inspired by [4]. As shown in Fig.3, given a new frame t+1, we first estimate the target location using a standard translation filter F_{tran}^{t}. Afterwards, we employ the learned scale filer F_{scale}^{t} at the target location to estimate the size of target. Finally, we update the F_{tran}^{t} and F_{scale}^{t}. Note that we only use the HOG feature to train the scale correlation filter F_{scale}^{t} after building the target scale pyramid while we employ the HOG and CN feature to estimate the target position by F_{tran}^{t}.

4. EXPERIMENTAL RESULTS

To evaluate our proposed tracker, we test it on the OTB50 datasets and compare it with 29 state-of-the-art trackers in the benchmark and 3 CFs based trackers. In this work, we employ the precision and success rate for quantitative analysis. In addition, we evaluate the speed of our tracker. Finally, we present 6 video sequences results for qualitative evaluation.

4.1. Quantitative Analysis

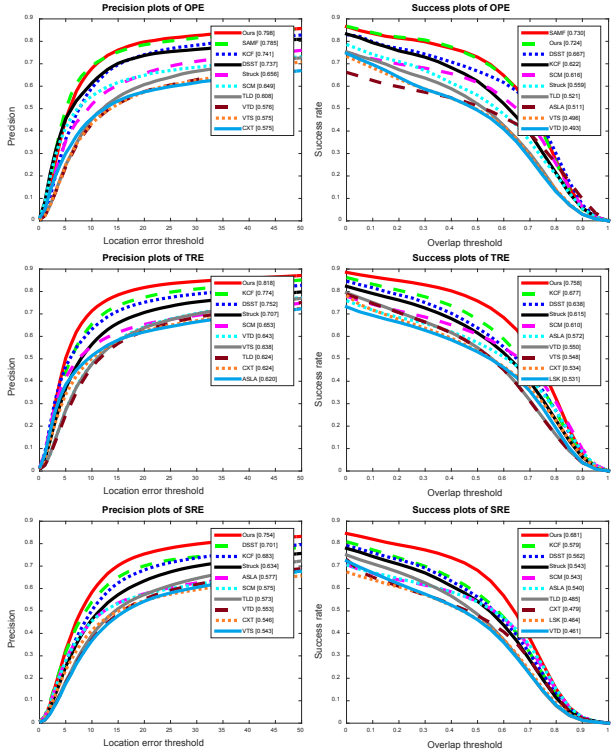


Fig.4. Plots of OPE, TRE and SRE on the OTB50. The performance scores for the top 10 trackers are shown in the legend.

To access the performance of the proposed tracker, two evaluation metrics, precision and success rate [6], are adopted in this paper. Precision rate refers to the percentage of frames where the estimated object location is within some threshold of the ground truth. In this work, we choose the threshold 20. Success rate is defined as $S = area(r_t \cap r_g)/area(r_t \cup r_g)$, where r_t denotes the tracked bounding box while r_g is the ground truth bounding box, and $\cap$ and $\cup$ denote the intersection and union of two regions respectively. To measure the performance, the area under curve (AUC) is employed to summarize and rank the performance of trackers. The precision plots and success plots of one-pass evaluation (OPE), temporal robustness evaluation (TRE) and spatial robustness evaluation (SRE) [6] are shown in Fig.4. We can see clearly that our proposed tracker is the best of all the trackers. In plots of OPE, our tracker performs well with precision rate of 79.8% and AUC score of 72.4% compared with other exiting state-of-the-art trackers. Our method obtains a gain of 5.7% in distance precision rate and 10.2% in overlap success rate, compared to our baseline tracker KCF.

4.2. Speed Analysis

In this section, we compare the speed of ours with several CF-based trackers on the same platform (Xeon E5-1607 3.0GHz, 8GB RAM). The average speed of our tracker on the OTB50 datasets is 22fps. The speed of CF-based trackers on several sequences are summarized in Table 1. From Table 1, we can see that our tracker is faster than the SAMF [5] but slower than the DSST [4] and KCF [3], and in Fig.4, our precision rates is 0.798, better than that of the SAMF 0.785. Therefore, our method achieves a good balance between efficiency and accuracy.

Table 1. The speed (fps) analysis of CFs based trackers.

	Basket ball	Bolt	Wal king	Wo man	Davi d2	Subw ay	Do g1
Ours	26	32	30	31	36	37	30
SAMF	7	9	7	10	24	16	18
DSST	28	37	35	36	58	55	42
KCF	261	371	343	290	461	495	335

4.3. Qualitative Analysis

To make qualitative comparisons, we plot the results of some challenging sequences for state-of-the-art trackers: ours (blue), KCF [3] (red), TLD [9] (green) and Struck [10] (black), shown in Fig.5. The KCF tracker performs well in deformation (Jogging1) and fast motion (CarScale) due to the robust HOG feature and dense sampling, but it drifts when heavy occlusions (Walking, Jogging1 and Tiger2) appears and it cannot deal with the scale change (CarScale, Walking2 and Dog1). In addition, the KCF tracker fails to handle motion blur (Jumping and Tiger2). The TLD tracker can re-detect objects in case of tracking failure, so it can track the car again at the frame 180 in the CarScale sequence. However, the TLD method does not fully exploit the target appearance and motion information as our approach. Therefore, it fails to track targets when targets undergo occlusions (Walking2 and Tiger2), deformation (Tiger2) and fast motion (CarScale and Tiger2). The Struck (black) performs badly in scale variance (CarScale and Walking2), occlusions (Walking2, Jogging1 and Tiger2) and fast motion (CarScale and Tiger2) since it is less effective in handling appearance change with one single classifier. Our tracker combines the target appearance and prior motion information to design a better response and uses mutiple features, so it is robust to occlusions (Waking2, Jogging1 and Tiger2), fast motion (CarScale and Tiger2) and deformation (Joggging2). In addition, our method can handle the scale change (CarScale and Walking2) due to its scale adaptive scheme.

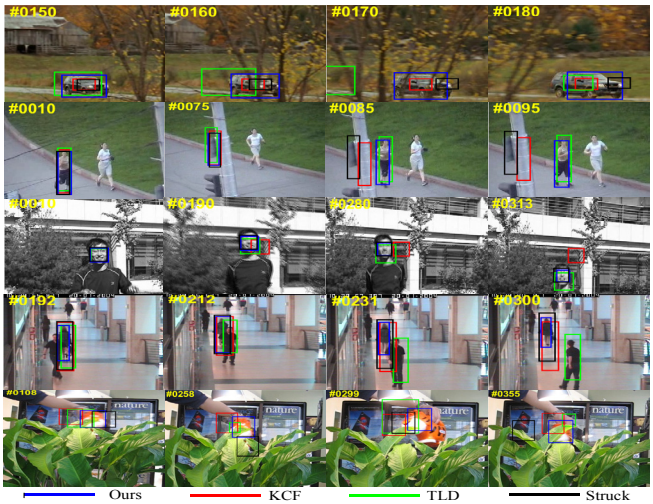


Fig.5. Comparison of our trackers (blue) with state-of-the-art trackers in challenging situations. From the top to bottom, the 5 video sequences are CarScale, Jogging1, Jumpping, Walking2 and Tiger2.

5. CONCLUSIONS

In this work, we develop a real-time target response adaptive and scale adaptive kernelized correlation filter tracker with multiple feature integration based on the framework of the KCF. Our tracker is robust to occlusion, fast motion, motion blur, scale variation and appearance changes. Extensive experiments validate the accuracy and efficiency of our tracker on the popular OTB50, which prove that our method achieves appealing results against state-of-the-art trackers in terms of accuracy and speed.

6. ACKNOWLEDGMENT

The Natural Science Foundation of Jiangsu Province supported this work (No. BK20150784), and the China Postdoctoral Science Foundation (No. 2015M581800), and the Fundamental Research Funds for the Central Universities (No. 30917011324).

7. REFERENCES

[1] Bolme D S, Beveridge J R, Draper B A, et al. Visual object tracking using adaptive correlation filters[C]. Computer vision and pattern recognition, 2010: 2544-2550.

[2] Henriques J F, Caseiro R, Martins P, et al. Exploiting the circulant structure of tracking-by-detection with kernels[C]. European conference on computer vision, 2012: 702-715.

[3] Henriques J F, Caseiro R, Martins P, et al. High-Speed Tracking with Kernelized Correlation Filters[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(3): 583-596.

[4] Danelljan M, Hager G, Khan F S, et al. Accurate Scale Estimation for Robust Visual Tracking[C]. British machine vision conference, 2014.

[5] Li Y, Zhu J. A Scale Adaptive Kernel Correlation Filter Tracker with Feature Integration[C]. European conference on computer vision, 2014: 254-265.

[6] Wu Y, Lim J, Yang M, et al. Online Object Tracking: A Benchmark[C]. Computer vision and pattern recognition, 2013: 2411-2418.

[7] Kristan M, Pflugfelder R, Leonardis A, et al. The visual object tracking VOT2014 challenge results[C]. European conference on computer vision, 2015: 191-217.

[8] Bibi A, Mueller M, Ghanem B. Target Response Adaptation for Correlation Filter Tracking[C]. European Conference on Computer Vision. 2016.

[9] Kalal Z, Matas J, Mikolajczyk K, et al. P-N learning: Bootstrapping binary classifiers by structural constraints[C]. Computer vision and pattern recognition, 2010: 49-56.

[10] Hare S, Saffari A, Torr P H, et al. Struck: Structured output tracking with kernels[C]. International conference on computer vision, 2011: 263-270.

[11] Ma C, Yang X, Zhang C, et al. Long-term correlation tracking[C]. Computer vision and pattern recognition, 2015: 5388-5396.

[12] Liu T, Wang G, Yang Q, et al. Real-time part-based visual tracking via adaptive correlation filters[C]. Computer vision and pattern recognition, 2015: 4902-4912.

[13] Bibi A, Ghanem B. Multi-template Scale-Adaptive Kernelized Correlation Filters[C]. International conference on computer vision, 2015: 613-620.

[14] De Weijer J V, Schmid C, Verbeek J, et al. Learning Color Names for Real-World Applications[J]. IEEE Transactions on Image Processing, 2009, 18(7): 1512-1523.

[15] Danelljan M, Hager G, Khan F S, et al. Discriminative Scale Space Tracking[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016: 1-1.

[16] Felzenszwalb P F, Girshick R, Mcallester D, et al. Object Detection with Discriminatively Trained Part-Based Models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(9): 1627-1645.

[17] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]. Computer vision and pattern recognition, 2005: 886-893.

[18] Atiya. Learning with kernels: Support vector machines, regularization, optimization, and beyond[J]. IEEE Transactions on Neural Networks, 2005, 16(3).

[19] Berlin B, Kay P. Basic color terms: their universality and evolution[J]. Language, 1999, 49(1).

[20] Khan F S, Anwer R M, De Weijer J V, et al. Color attributes for object detection[C]. computer vision and pattern recognition, 2012: 3306-3313.

[21] Khan F S, De Weijer J V, Vanrell M, et al. Modulating Shape Features by Color Attention for Object Recognition[J]. International Journal of Computer Vision, 2012, 98(1): 49-64.

[22] Khan F S, Anwer R M, De Weijer J V, et al. Coloring Action Recognition in Still Images[J]. International Journal of Computer Vision, 2013, 105(3): 205-221.