Detecting Image Frames which Contain a Moving Object from a Severely Distorted Video Stream Using Dynamic Mode Decomposition

Daisuke Uchida[a], Atsushi Yamashita[a] and Hajime Asama[a]

[a]The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan

ABSTRACT

There are few moving object detection techniques dealing with severely distorted video imagery, such as one taken from above the wavy water surface. In this paper, a method that identifies image frames containing a moving object from a video taken from above the wavy water surface is proposed. Considering the difficulty to apply common video processing techniques to such a video suffering from the severe distortion, the proposed method utilizes dynamic mode decomposition, a data-driven method for analysis of dynamical systems, to develop an algorithm that extracts information of a moving object from a video stream. The experimental evaluation shows that the proposed method is able to identify image frames containing a moving object from a severely distorted video stream.

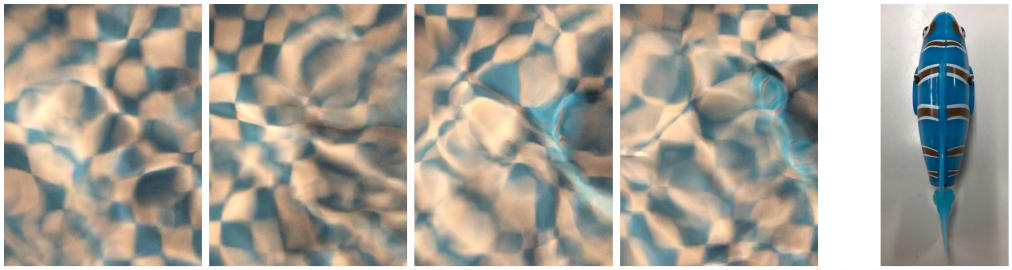
Keywords: Moving object detection, imaging through water-air interface, dynamic mode decomposition

1. INTRODUCTION

Rivers are challenging environments in which to study fish since flowing water and variable depths make rivers not only difficult to sample but also dangerous.[1] Underwater video analysis is one of common technological solutions to study fish in rivers, and several works dealt with moving object detection, tracking, and classification for underwater videos.[2–6] However, the use of underwater imaging may be limited in several situations, for example, the headwater stream where it is difficult to set up underwater cameras due to rapid flow and/or no enough space. In such a case, imaging from above the water surface is considered to be a reasonable solution, while videos taken from above the water flow suffer from the severe distortion through the water-air interface and make it difficult to apply video processing techniques (see Fig. 1). For example, one of the most fundarmental techniques for video analysis is moving object detection, and several algorithms for background subtraction have been developed.[7–9] However, it is difficult to model the appropriate background model in this case, since the heavy distortion due to the wavy water surface leads to continuous and drastic changes of the intensities over the frame, which make it extremely difficult to distinguish the foreground from the background properly.

In this paper, moving object detection is considered and a method to detect image frames which contain a moving object from a severely distorted video taken from above the water surface is developed. Although several studies dealt with image reconstruction from image frames suffering from the distortion due to the wavy water sueface,[10–13] they need to assume static scenes. On the other hand, a moving object detection method for imaging through water-air interface has been developed.[14] However, it considered relatively calm waves, while this paper deals with highly complicated shapes of waves such as ones in the headwater stream, which result in the heavily distorted imagery where most information of objects' shapes is lost (see frame 286 and 287 in Fig. 1).

The proposed method takes the advantage of dynamic mode decomposition (DMD), a data-driven method for analysis of dynamical systems,[15–18] and extracts information of moving objects as the specific DMD mode from a recorded video.



frame 38. frame 198. frame 286. frame 287. Rigid body robotic fish.

Figure 1. Frames taken through wavy water surface in an experiment and a rigid body robitic fish (frame 286 and 287 contain a robotic fish as a moving object).

2. DYNAMIC MODE DECOMPOSITION

Dynamic mode decomposition (DMD) is a data-driven method for analyzing nonlinear dynamical systems originally introduced in the fluid mechanics community.[15] DMD has been applied to not only the fluid mechanics but also other several research fields, including image processing and computer vision.[19–21] Given high-dimensional data in the form of two matrices, DMD decomposes the data into a sum of several modes called DMD modes. In this section, the formulation of DMD is described within the context of the time series data $\{\boldsymbol{x}_i \in \mathbb{R}^N\}_{i=0}^M$ evenly spaced by Δt, while DMD can be defined in more generalized conditions.[16]

Two matrices $\boldsymbol{X}$ and $\boldsymbol{X}'$ are formed as follows:

$$\boldsymbol{X} = [\boldsymbol{x}_0 \ \cdots \ \boldsymbol{x}_{M-1}], \quad \boldsymbol{X}' = [\boldsymbol{x}_1 \ \cdots \ \boldsymbol{x}_M]. \tag{1}$$

Then DMD modes are defined as the eigenvectors of a matrix $\boldsymbol{A}$, which is defined as:

$$\boldsymbol{A} := \boldsymbol{X}'\boldsymbol{X}^+ \in \mathbb{R}^{N\times N}, \tag{2}$$

where $\boldsymbol{X}^+$ denotes the Moore-Penrose pseudoinverse of $\boldsymbol{X}$. For high-dimensional data such as video streams, the size of $\boldsymbol{A}$ may be quite large. Hence a lower-dimensional matrix $\tilde{\boldsymbol{A}}$ are introduced using the POD modes of $\boldsymbol{A}$. First, take the singular value decomposition of $\boldsymbol{X}$:

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T. \quad \left(\boldsymbol{U} \in \mathbb{R}^{N\times r}, \quad \boldsymbol{\Sigma} \in \mathbb{R}^{r\times r}, \quad \boldsymbol{V} \in \mathbb{R}^{M\times r}, \quad \mathrm{rank}(\boldsymbol{X}) = r\right) \tag{3}$$

Then the matrix $\tilde{\boldsymbol{A}}$ are defined as follows:

$$\tilde{\boldsymbol{A}} := \boldsymbol{U}^T\boldsymbol{A}\boldsymbol{U} = \boldsymbol{U}^T\boldsymbol{X}'\boldsymbol{V}\boldsymbol{\Sigma}^{-1}. \tag{4}$$

The eigendecomposition of $\tilde{\boldsymbol{A}}$ is described using a matrix $\boldsymbol{W}$ and a diagonal matrix $\boldsymbol{\Lambda}$:

$$\tilde{\boldsymbol{A}}\boldsymbol{W} = \boldsymbol{W}\boldsymbol{\Lambda}, \quad (\boldsymbol{W} := [\,\boldsymbol{w}_1 \ \cdots \ \boldsymbol{w}_r\,], \quad \boldsymbol{\Lambda} := \mathrm{diag}(\lambda_1, \cdots, \lambda_r)) \tag{5}$$

where $(\boldsymbol{w}_i, \lambda_i)$ is a pair of i-th eigenvector/eienvalue of $\tilde{\boldsymbol{A}}$. Finally, the DMD modes $\{\boldsymbol{\varphi}_i\}_{i=1}^r$ are obtained as follows:

$$\boldsymbol{\varphi}_i = \boldsymbol{X}'\boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{w}_i. \tag{6}$$

It was shown that the DMD modes defined above are exact eigenvectors of $\boldsymbol{A}$ provided $\lambda_i \neq 0$, and conversely, all nonzero eigenvalues and their corresponding eigenvectors of $\boldsymbol{A}$ are given by $(\boldsymbol{\varphi}_i, \lambda_i)_{i=1}^r$.[16]

From eq. (2) it is clear that the DMD algorithm seeks the best-fit linear operator which defines a discrete-time linear dynamical system:

$$\boldsymbol{x}_{i+1} = \boldsymbol{A}\boldsymbol{x}_i, \quad (i = 0, \cdots, M-1) \tag{7}$$

and the relationship (7) holds if $\mathrm{null}(\boldsymbol{X}) \subset \mathrm{null}(\boldsymbol{X}')$.[16] When the data $\{\boldsymbol{x}_i \in \mathbb{R}^N\}_{i=0}^M$ are sampled from a continuous-time system, corresponding eigenvalues of this system can be shown to be

$$\omega_i = \frac{\ln \lambda_i}{\Delta t}, \tag{8}$$

Further author information: (Send correspondence to Daisuke Uchida) E-mail: uchida@robot.t.u-tokyo.ac.jp

and the DMD reconstruction of the data at time t is given by

$$\boldsymbol{x}_{\mathrm{DMD}} = \sum_{i=1}^{r} \exp(\omega_i t)\, b_i \boldsymbol{\varphi}_i, \tag{9}$$

where

$$[\, b_1 \, \cdots \, b_r \,]^T := [\, \boldsymbol{\varphi}_1 \, \cdots \, \boldsymbol{\varphi}_r \,]^{+} \boldsymbol{x}_0. \tag{10}$$

3. PROPOSED ALGORITHM

3.1 Analysis of DMD modes of severely distorted video streams

In the proposed algorithm, DMD modes of video streams are sorted by their frequencies $\{|\mathrm{Im}(\omega_i)|\}$. Specifically, a DMD mode which has the lowest frequency is named first DMD mode, a DMD mode which has the second lowest frequency is named second DMD mode, and so on. If several DMD modes have the same frequency in common with each other, the mode names are determined by their decay/growth rates $\{|\mathrm{Re}(\omega_i)|\}$, which are sorted from smallest to largest. Therefore, the first DMD mode denotes one with the eigenvalue nearest the origin.

Objects in recorded videos are classified into two categories, foreground and background. Here we define the foreground as a moving object in the scene, which is our interest. Hence, the background denotes all static objects in the scene, or the static scene. Moreover, actual recorded video streams consist of three parts, the foreground, the background, and the distortion due to the wavy water-air interface. Note that only one moving object is considered for simplicity.

It is clear that the background corresponds to the first DMD mode, whose eigenvalue is ideally located at the origin, that is, constant intensities of the background correspond to the DC component of the DMD modes. Furthermore, the proposed algorithm assumes that the foreground and the distortion correspond to the second DMD mode and other high-frequency DMD modes, respectively. The foreground may correspond to the second DMD mode, whose frequency is next lowest to the background mode, since complicated and fine shapes of the water surface which distort imagery so heavily that most information of the scene is lost (see Fig. 1) may be converted into high-frequency signals. Although waves which have relatively large wavelengths may be converted into low-frequency signals, they do not result in the heavy distortion as shown in Fig. 1, which is our main concern. Thus, we assume the signals of the distortion which have higher frequencies than that of the foreground, or the second DMD mode.

On the other hand, if a video stream contains no moving object, it consists of only two parts, the background and the distortion, which means that the second DMD mode in this case also corresponds to the distortion by definition. Namely, a physical meaning of the second DMD mode depends on if the video contains a moving object or not, and this observation is utilized to evaluate the spatial patterns of DMD modes in the sequel.

3.2 Evaluation of the spatial patterns of DMD modes

In this section, the spatial patterns of second DMD modes are examined. From a second DMD mode $\boldsymbol{\varphi}_2$, the proposed algorithm generates a binary image $\boldsymbol{P} \in \mathbb{R}^{n \times m}$ which is given by:

$$(\boldsymbol{P})_{ij} = \begin{cases} 1 & (\mathrm{abs}[(\boldsymbol{\varphi}_2)_{ij}] \geq p_{\mathrm{th}}) \\ 0 & (\mathrm{others}) \end{cases}, \quad (1 \leq i \leq n,\ 1 \leq j \leq m) \tag{11}$$

where p_{th} denotes a threshold. Patterns in $\boldsymbol{P}$ show the features of the corresponding DMD modes, or their counterparts (the background, the foreground, and the distortion) if the threshold p_{th} is chosen suitably. Hence, if there is no moving object in a video stream, $\boldsymbol{P}$ represents the patterns of the distortion all over the frame, while if a video stream contains a moving object, it shows the shape and/or trajectory of it.

This difference is evaluated by the area of the patterns in $\boldsymbol{P}$. Considering that the patterns of a moving object may appear in relatively large size in the frame compared with those of the distortion which is caused by

the fine water surface waves, $\boldsymbol{P}$ is first split into a set of patches $\{\boldsymbol{P'}(s) \in \mathbb{R}^{n' \times m'}\}_{s=1}^{l}$. Then a value J', which corresponds to the area of the patterns of a moving object, is calculated as follows:

$$J' = \begin{cases} 1 & \left(\sum_{i=1}^{n'} \sum_{j=1}^{m'} (\boldsymbol{P'}(s))_{ij} = m'n'\right) \\ 0 & \text{(others)} \end{cases} . \tag{12}$$

Note that the patterns of the distortion will result in $J' = 0$ since they tend to be small and fine in the frame. Finally, the J's are added up to calculate a value called score J, which is expected to take higher values when the frames contain a moving object than when they contain no moving component in the scene.

3.3 Determination of the threshold

In the evaluation of the spatial patterns of the second DMD modes, the representation of the patterns in $\boldsymbol{P}$ is sensitive to the value of the threshold p_{th}. For example, Fig. 2 shows patterns in $\boldsymbol{P}$ generated from a recorded video taken in an experiment, which is explained in section 4. The frames corresponding to (a)/(b) are shown in Fig. 3, and the frames corresponding to (b) contain a moving object. As shown Fig. 2, if the threshold is set to the appropriate value ($p_{\text{th}} = 0.002$ in this case), (a) and (b) represent fine patterns of the distortion over the frame and relatively large patterns of the shape or trajectory corresponding to the moving object on the upper right of the frame, respectively. However, if the threshold is smaller ($p_{\text{th}} = 0.0005$) or larger ($p_{\text{th}} = 0.003$), the difference of the patterns disappears, and J' introduced in the previous section tends to get the same value.

Considering this sensitivity, the proposed algorithm first sets p_{th} to be quite large which results in $\boldsymbol{P} = \boldsymbol{0}$, and generates multiple $\boldsymbol{P}$s making the value of p_{th} δp smaller at one iteration. It stops generating $\boldsymbol{P}$s when patterns appear in more than half of the frame, since under the condition that $p_{\text{th}} \ll 1$, the score J gets a high value even if the scene contains no moving object. Note that this condition is confirmed using another set of patches $\{\boldsymbol{P''}(i) \in \mathbb{R}^{n'' \times m''}\}_{i=1}^{q}$. The proposed algorithm is summarized in Algorithm 1.

4. EXPERIMENTAL EVALUATION

4.1 Experimental set up

An experiment was conducted in order to confirm the effectiveness of the proposed algorithm. A video stream was recorded from above a tank filled with water in a laboratory, where a water pump set to near the water surface generates waves. A rigid body robotic fish, which is shown in Fig. 1 was used to simulate a moving object (a fish) in the headwater stream. Figure 1 also shows some sample frames taken in the experiment.

4.2 Detection of frames which contain a moving object

The score J in the algorithm 1 is shown in Fig. 3. The period where a moving object is in the frame is colored in the figure. The frame size was $(n, m) = (1080\,\text{pixel}, 820\,\text{pixel})$, and the sizes of the two patches are set to $(n', m') = (100\,\text{pixel}, 100\,\text{pixel})$ and $(n'', m'') = (54\,\text{pixel}, 41\,\text{pixel})$,respectively. Note that the patches $\{\boldsymbol{P'}(s)\}$ were generated with a 90 % overlap between each pair of adjacent patches. 30 frames ($M = 30$, 1 s) were used as the input data to DMD at a time, and δp was set to 0.0005.

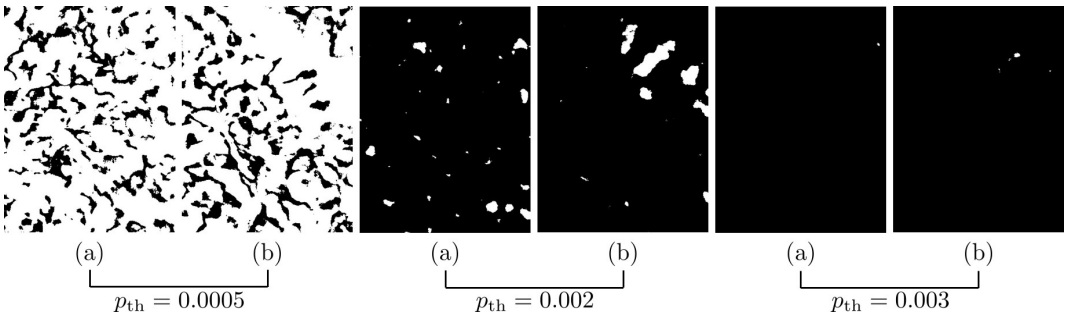


Figure 2. Patterns in $\boldsymbol{P}$ (The frames corresponding to (a)/(b) are shown in Fig. 3).

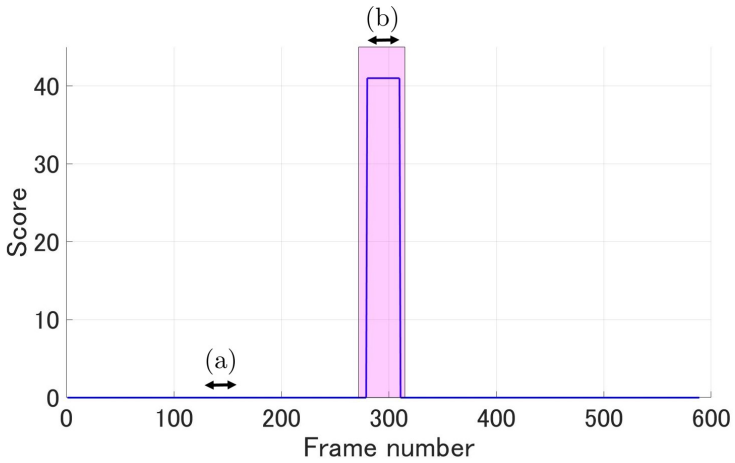


Figure 3. Score J (frames containing a moving object are colored).

The score J got high values at the frames containing a moving object, whereas it got 0 at other frames, which capture the static scene. Therefore, it is able to identify the frames which contains a moving object from the time series of the score J.

5. CONCLUSION

In this paper, an algorithm which detects the frames which contain a moving object from a severely distorted video stream that is taken through wavy water surface was proposed. By analyzing the DMD modes of the video

Algorithm 1 Proposed algorithm

Input: $\{\boldsymbol{I}(i) \in \mathbb{R}^{n\times m}\}_{i=1}^{H}$
Output: $J = J(k)$
  for $k = 1 : \lfloor \frac{H-1}{M+1} \rfloor$ **do**
    $J(k) = 0$
    do
      $\boldsymbol{\varphi}_2$ =DMD($\{\boldsymbol{I}(i)\}_{i=k}^{k+M-1}$)
      $(\boldsymbol{P})_{ij} = \begin{cases} 1 & (\mathrm{abs}(\boldsymbol{\varphi}_2)_{ij} \geq p_{\mathrm{th}}) \\ 0 & (\mathrm{others}) \end{cases} \qquad (1 \leq i \leq n,\ 1 \leq j \leq m)$
      $p_{\mathrm{th}} \leftarrow p_{\mathrm{th}} - \delta p$
      Split $\boldsymbol{P}$ into two sets of patches: $\{\boldsymbol{P'}(i) \in \mathbb{R}^{n'\times m'}\}_{i=1}^{l}$, $\{\boldsymbol{P''}(i) \in \mathbb{R}^{n''\times m''}\}_{i=1}^{q}$
      for $s = 1 : l$ **do**
        $J' = \begin{cases} 1 & \left(\sum_{i=1}^{n'} \sum_{j=1}^{m'} (\boldsymbol{P'}(s))_{ij} = m'n'\right) \\ 0 & (\mathrm{others}) \end{cases}$
        $J(k) \leftarrow J(k) + J'$
      end for
      $e = 0$
      for $s = 1 : q$ **do**
        $e' = \begin{cases} 1 & \left(\sum_{i=1}^{n''} \sum_{j=1}^{m''} (\boldsymbol{P''}(s))_{ij} \neq 0\right) \\ 0 & (\mathrm{others}) \end{cases}$
        $e \leftarrow e + e'$
      end for
    while $e/q < 1/2$
    $k \leftarrow k + M$
  end for

stream, the proposed algorithm evaluated the score J as the output, which was expected to take higher values when a moving object is in the frame than when it is not. An experimental evaluation showed that the frames which contain a moving object was able to be identified from the time series of the J.

REFERENCES

[1] Cooke, S., Paukert, C., and Hogan, Z., "Endangered river fish: factors hindering conservation and restoration," *Endangered Species Research* **17**, 179–191 (2012).

[2] Forczmański, P., Nowosielski, A., and Marczeski, P., "Video stream analysis for fish detection and classification," *Advances in Intelligent Systems and Computing* **342**, 157–169 (2015).

[3] Spampinato, C., Chen-Burger, Y.-H., Nadarajan, G., and Fisher, R., "Detecting, tracking and counting fish in low quality unconstrained underwater videos," in [*Proceedings of the Third International Conference on Computer Vision Theory and Applications*], **2**, 514–519 (2008).

[4] Shevchenko, V., Eerola, T., and Kaarna, A., "Fish detection from low visibility underwater videos," in [*Proceedings of the 2018 24th International Conference on Pattern Recognition*], 1971–1976 (2018).

[5] Evans, F., "Detecting fish in underwater video using the EM algorithm," in [*Proceedings of the 2003 International Conference on Image Processing*], III–1029 (2003).

[6] Zion, B., Shklyar, A., and Karplus, I., "In-vivo fish sorting by computer vision," *Aquacultural Engineering* **22**(3), 165–179 (2000).

[7] Zivkovic, Z. and van der Heijden, F., "Efficient adaptive density estimation per image pixel for the task of background subtraction," *Pattern Recognition Letters* **27**(7), 773–780 (2006).

[8] Barnich, O. and Van Droogenbroeck, M., "ViBe: A universal background subtraction algorithm for video sequences," *IEEE Transactions on Image Processing* **20**(6), 1709–1724 (2011).

[9] Zivkovic, Z., "Improved adaptive gaussian mixture model for background subtraction," in [*Proceedings of the 17th International Conference on Pattern Recognition*], **2**, 28–31 (2004).

[10] Efros, A., Isler, V., Shi, J., and Visontai, M., "Seeing through water," in [*Proceedings of the Neural Information Processing Systems Conference*], (2004).

[11] Oreifej, O., Shu, G., Pace, T., and Shah, M., "A two-stage reconstruction approach for seeing through water," in [*Proceedings of the Conference on Computer Vision and Pattern Recognition 2011*], 1153–1160 (2011).

[12] Tian, Y. and Narasimhan, S., "Seeing through water: Image restoration using model-based tracking," in [*Proceedings of the 2009 IEEE 12th International Conference on Computer Vision*], 2303–2310 (2009).

[13] Zhang, Z. and Yang, X., "Reconstruction of distorted underwater images using robust registration," *Optics Express* **27**, 9996–10008 (2019).

[14] Alterman, M., Schechner, Y., Perona, P., and Shamir, J., "Detecting motion through dynamic refraction," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35**(1), 245–251 (2013).

[15] Schmid, P. and Sesterhenn, J., "Dynamic mode decomposition of numerical and experimental data," in [*Sixty-First Annual Meeting of the APS Division of Fluid Dynamics*], (2008).

[16] Tu, J., Rowley, C., Luchtenburg, D., Brunton, S., and Kutz, N., "On dynamic mode decomposition: Theory and applications," *Journal of Computational Dynamics* **1**(2), 391–421 (2014).

[17] Rowley, C., Mezić, I., Bagheri, S., Schlatter, P., and Henningson, D., "Spectral analysis of nonlinear flows," *Journal of Fluid Mechanics* **641**, 115–127 (2009).

[18] Schmid, P., "Dynamic mode decomposition of numerical and experimental data," *Journal of Fluid Mechanics* **656**, 5–28 (2010).

[19] Kutz, J. N., Fu, X., Brunton, S. L., and Erichson, N. B., "Multi-resolution dynamic mode decomposition for foreground/background separation and object tracking," in [*Proceedings of the 2015 IEEE International Conference on Computer Vision Workshop*], 921–929 (2015).

[20] Sikha, O. K. and Soman, K. P., "Multi-resolution dynamic mode decomposition-based salient region detection in noisy images," *Signal, Image and Video Processing* , 1–9 (2019).

[21] Le Ngo, A. C., Liong, S., See, J., and Phan, R. C., "Are subtle expressions too sparse to recognize?," in [*Proceedings of the 2015 IEEE International Conference on Digital Signal Processing*], 1246–1250 (2015).