Robust and Fast
Eulerian Video Magnification
for Practical Applications

March 2021

Shoichiro Takeda

Robust and Fast
Eulerian Video Magnification
for Practical Applications

Shoichiro Takeda

Graduate School of Science and Technology

Degree Programs in Systems and Information Engineering

University of Tsukuba

March 2021

Abstract

In our world, there are many kinds of subtle yet important physical/natural phenomena. For example, human skin color subtly changes with blood circulation caused by heart beating, drones fluctuate slightly for stabilizing themselves during flight, and subtle vibrations of buildings reflect their structural stability. Therefore, such subtle color changes or motions are essential for practical applications where you need to understand scene contexts or anomalous behavior correctly. However, they are often difficult to see with human eyes.

For revealing such subtle changes, Eulerian video magnification (EVM) methods have been proposed. The EVM methods try to amplify only subtle changes in a video and then reveal them for users via the synthesis amplification video result. However, conventional EVM methods have the following three problems for practical applications where only subtle changes caused by physical/natural phenomena need to be revealed quickly and correctly. (Problem 1) A promising EVM method, called the Eulerian video acceleration magnification (EVAM) method, can ignore only slow large motions of objects when revealing subtle changes in a video. However, the EVAM method cannot ignore quick large motions and thus produces messy artifacts when objects in a video move quickly and largely. (Problem 2) Subtle changes in a video contain meaningful ones caused by physical/natural phenomena and non-meaningful ones caused by photographic subtle noise. Some conventional EVM methods address to detect only the meaningful subtle changes but are insufficient and have severe limitations. Thus, conventional EVM methods often mistakenly amplify photographic subtle noise and produce noisy and/or misleading results. (Problem 3) Conventional EVM methods construct over-complete image pyramid representations when analyzing subtle changes in a video, and thus require a long computational time in proportion to video resolution and time frame length.

This dissertation is comprised of three EVM studies which focus on overcoming the above three EVM problems (Problems 1, 2, and 3), in order to facilitate the robust and fast analysis of subtle changes in a video and enhance the performance of EVM for practical applications.

First, we committed an EVM study for ignoring large motions of objects and revealing only subtle changes in a video. For this purpose, we used a differential feature called jerk to make the EVAM method robust even to quick large motions as well as slow ones. We showed that our method produces impressive EVM results without messy artifacts, which could be caused by slow and/or quick large motions of objects, in both real and synthetic videos.

Second, we committed an EVM study for ignoring photographic subtle noise and revealing only the meaningful subtle changes in a video. For this purpose, we proposed an EVM method using both a fractional anisotropy and edge-aware regularization. The use of them can effectively suppress the effect of photographic subtle noise, and thus our method can ignore photographic subtle noise and produces impressive EVM results in both real and synthetic videos.

Third, we committed an EVM study for accelerating computational time of EVM. For this purpose, on the basis of signal correlation between adjacent pyramid levels, we constructed fewer image pyramid representations than the original ones when analyzing subtle changes in a video. We showed that our method produces impressive EVM results equivalent to conventional ones within a short computational time in both real and synthetic videos.

With the above three EVM studies, subtle yet important physical/natural phenomena can be quickly and correctly revealed even under the practical conditions, where large motions of objects exist (Problem 1), photographic subtle noise in a video exist (Problem 2), and short computational time is required (Problem 3). Finally, we conclude this dissertation by clarifying our contributions and future work

to enhance the performance of EVM for practical applications.

Acknowledgements

First and foremost, I would like to thank my supervisor, Prof. Yoshihiro Kanamori. This dissertation has been done under his direction, and would not have been possible without his helpful advice, broad knowledge for computer science, and constant encouragement. I am also appreciated to Prof. Yuki Endo, who gave insightful comments and suggestions to this dissertation.

I would also like to thank my ex-supervisors, Prof. Junichi Ushiba, Prof. Junichi Ushiyama, and Prof. Shoko Kasuga, during bachelor's and master's programs at Keio University. Their constant encouragement, broad knowledge for neuroscience, and paper writing expertise allowed me to grow as a research scientist. I learned visions and philosophy of research from them.

In addition, I would like to express my gratitude to my colleagues of Nippon Telegraph and Telephone (NTT) Corporation for allowing me to grow as a researcher and person. I would like to especially thank the members at NTT Media Intelligence Laboratories, Dr. Hideaki Kimata, Dr. Megumi Isogai, Dr. Dan Mikami, Dr. Shinya Shimizu, Kazuki Okami, Akio Kameda, Masaaki Matsumura, and the member at NTT Service Evolution Laboratories, Yasunori Akagi.

Furthermore, I would like to thank the members of Computational Geometry and Graphics Lab in University of Tsukuba for giving me constructive comments and warm encouragement.

I am grateful to the great faculty members, Prof. Jun Mitani, Prof. Kazuhiro Fukui, Prof. Keisuke Kameyama, and Prof. Makoto Okabe, for reviewing this dissertation and giving insightful comments and suggestions.

Finally, I would like to thank my family and friends for always being supportive.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

With the recent advances in camera technology, many kinds of physical/natural phenomena can now be easily captured over different space and time scales. However, we human often fail to visually perceive such phenomena if they are extremely small. For example, human skin color subtly changes with blood circulation caused by heart beating. These subtle color changes are too small to see with human eyes but can be the clues to extract pulse rate [1, 2, 3]. Similarly, subtle motions, hard for humans to see, are often used for evaluating or comparing some events, e.g., tiny vibrations of strings in instruments play wonderful sounds, high-quality drones subtly sway to make themselves stabilize in quick flight. Therefore, such subtle color changes or motions are essential for practical applications where you need to understand scene contexts or anomalous behavior correctly. However, again, they are difficult to see with human eyes.

For revealing such subtle changes, Eulerian video magnification (EVM) methods have been proposed to amplify subtle changes in a video [4, 5, 6, 7, 8, 9]. These EVM methods are based on Eulerian description that measures subtle color changes or motions of objects in a video as subtle signals (e.g., color signals or phase signals representing local motions [10]) over time frames at each pixel position. In these EVM methods, the color/phase signals at each pixel position are temporally bandpass filtered with a target temporal frequency. The bandpass signals are then amplified with an amplification factor to reveal subtle changes in a video. However, conventional EVM methods have the following three problems for practical applications where only subtle changes caused by physical/natural phenomena need to be revealed quickly and correctly.

(Problem 1) Large Motions in Video. A promising EVM method, called the Eulerian video acceleration magnification (EVAM) method [9], can ignore only slow large motions of objects when revealing subtle changes in a video. However, the EVAM method cannot ignore quick large motions and thus produces messy artifacts when objects in a video move quickly and largely.

(Problem 2) Photographic Subtle Noise in Video. Subtle changes in a video contain meaningful ones caused by physical/natural phenomena and non-meaningful ones caused by photographic subtle noise. Some conventional EVM methods address to detect only the meaningful subtle changes but are insufficient and have severe limitations. Thus, conventional EVM methods often mistakenly amplify photographic subtle noise and produce noisy and/or misleading results.

(Problem 3) Long Computational Time. Conventional EVM methods construct over-complete image pyramid representations when analyzing subtle changes in a video, and thus require a long computational time in proportion to video resolution and time frame length.

This dissertation is comprised of three EVM studies which focus on overcoming the above three EVM problems (Problems 1, 2, and 3), in order to facilitate the robust and fast analysis of subtle changes in a video and enhance the performance of EVM for practical applications. To overcome these problems, we kept in mind to propose simple yet novel solutions. Specifically, we tackled problems 1 and 2 by simple spatio-temporal filtering techniques (without the necessary of complex optimization procedure required in, e.g., deep learning techniques) utilizing the knowledge of neuroscience to which I belonged until my bachelor's and master's programs. We consider that this knowledge utilization is the most interesting point in this dissertation because it solves computer science problem from a completely different research perspective. Here, brief overviews of the three EVM studies are described as follows.

(Solution 1) Ignoring Large Motions in Video. In Chapter 4, we committed an EVM study for ignoring large motions of objects and revealing only subtle changes in a video. For this purpose, we proposed an EVM method that combines the Eulerian video acceleration magnification (EVAM) method [9], which ignores only slow large motions, with a differential feature called jerk. This method is consisted of making the EVAM method robust even to quick large motions as well as slow ones by utilizing jerk. Jerk has been used to evaluate smoothness of time series data in neuroscience [11, 12, 13] and can be used to identify steep changes in the color/phase signals caused by quick large motions of objects. We showed that our method produces impressive EVM results without messy artifacts, which could be caused by slow and/or quick large motions of objects, in both real and synthetic videos.

(Solution 2) Ignoring Photographic Subtle Noise in Video. In Chapter 5, we committed an EVM study for ignoring photographic subtle noise and revealing only the meaningful subtle changes in a video. For this purpose, we propose an EVM method using both a fractional anisotropy (FA) and edge-aware regularization. FA has been used in neuroscience to evaluate anisotropic diffusion of water molecules in the body for revealing the shape of tiny nerve cells [14, 15], and we thus consider that FA can identify anisotropic temporal diffusion of the meaningful subtle changes caused by physical/natural phenomena. Additionally, the edge-aware regularization can refine uncertain subtle motions at flat (texture-less) regions in a video. We showed that our method effectively ignores photographic subtle noise compared with conventional EVM methods and produces impressive EVM results in both real and synthetic videos.

(Solution 3) Accelerating Computational Time. In Chapter 6, we committed an EVM study for accelerating computational time of EVM. For this purpose, we propose an EVM method that combines local image processing with a conventional fast EVM method with the Riesz pyramid [6]. On the basis of signal correlation between adjacent pyramid levels as reported in [16, 17], our method constructs fewer image pyramid representations than those constructed in the Riesz method when analyzing subtle changes in a video. We showed that our method produces impressive EVM results equivalent to conventional ones within a short computational time in both real and synthetic videos.

This dissertation consists of seven chapters. In Chapter 2, we explain the major related work of EVM and video color/motion analysis for practical applications. In Chapter 3, we introduce preliminary formulation of EVM to clearly explain the following chapters. In Chapters 4, 5, and 6, we explain the details of each study (Solutions 1, 2, and 3). Finally, Chapter 7 summarizes the conclusions and contributions of this dissertation and provides future work directions.

1.1 Publications and Awards

1.1.1 Reference Publications

This dissertation largely refers to the following three publications.

International Conference Proceedings (with Peer Review)

1. Shoichiro Takeda, Kazuki Okami, Megumi Isogai, Dan Mikami, and Hideaki Kimata: "Jerk-Aware Video Acceleration Magnification," *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.1769-1777, 2018.
2. Shoichiro Takeda, Yasunori Akagi, Kazuki Okami, Megumi Isogai, and Hideaki Kimata: "Video magnification in the wild using fractional anisotropy in temporal distribution," *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.1614-1622, 2019.

Journal Papers (with Peer Review)

1. Shoichiro Takeda, Megumi Isogai, Shinya Shimizu, and Hideaki Kimata: "Local Riesz pyramids for faster phase-based video magnification," *IEICE Transactions on Information and Systems*, Vol.E103-D, No.10, pp.2036-2046.

1.1.2 Other Publications

The following publications are also supplementary referred in this dissertation.

Domestic Conference Proceedings (without Peer Review)

1. Shoichiro Takeda, Megumi Isogai, and Hideaki Kimata: "Jerk-Aware Filter for Video Magnification," *Image Media Processing Symposium (IMPS)*, 2017 (in Japanese).

2. Shoichiro Takeda, Kazuki Okami, Masaaki Matsumura, Megumi Isogai, and Hideaki Kimata: "Jerk-Aware Video Magnification – Characterization of Jerk-Aware Filter –," *The Meeting on Image Recognition and Understanding (MIRU)*, 2018 (in Japanese).

3. Shoichiro Takeda, Akio Kameda, Megumi Isogai, and Hideaki Kimata: "A Study of Quality of Experience Assessment for Video Magnification," *IPSJ Special Interest Group on Audio Visual and Multimedia Information Processing (IPSJ-AVM)*, 2019 (in Japanese).

4. Shoichiro Takeda, Megumi Isogai, Shinya Simizu and Hideaki Kimata: "Local Riesz Pyramids for Faster Phase-based Video Magnification," *IPSJ Special Interest Groups on Computer Graphics and Visual Informatics (IPSJ-CGVI)*, 2019 (in Japanese).

1.1.3 Related Awards

The following awards are related to this dissertation.

1. February 2019: AVM Award from IPSJ Special Interest Group on Audio Visual and Multimedia Information Processing.

2. February 2019: Excellent Research and Presentation Award from IPSJ Special Interest Groups on Computer Graphics and Visual Informatics.

3. December 2019: Research and Development Encouragement Award from NTT Service Innovation Laboratory Group.

Chapter 2

Related Work

In this chapter, we explain the major related work of EVM. We also explain the related interesting work in a video color/motion manipulation/analysis that handle temporal deviations in a video as the EVM research does.

2.1 Lagrangian Video Motion Magnification

The first video motion magnification method is proposed by Liu et al. [18] with Lagrangian description that measures motions in a video by matching feature points between frames and estimating optical flow. Liu et al. [18] measure motions of objects by using optical flow analysis [19] to segment background motions and subtle motions of interest for magnification in a video by using energy minimization estimation that involves with motion likelihood, color likelihood, and spatial connectivity [20]. After the segmentation, they magnify only the subtle motions of interest. Moreover, since the motion magnification will reveal occluded regions, texture inpainting [21] needs to be applied to those regions as a post-processing. The motion magnification results amazed the world because they first show that imperceptible motions in a video can be revealed with the computer vision/graphics technology.

However, this Lagrangian-based method is computationally expensive because it is consisted of a complex combination of three algorithms, namely, optical flow analysis [19], segmentation [20], and texture in-painting [21]. Additionally, the each algorithm has been still researched as an unsolved problem [22, 23, 24, 25, 26].

2.2 Learning-based Video Motion Magnification

Recently, learning-based video motion magnification (LebVMM) methods with a convolutional neural network (CNN) have been proposed [27, 28]. Oh et al. [27] newly design an encoder-decoder CNN motion magnification model that is trained by supervised learning with synthetic training dataset created by the authors. The trained encoder CNN model takes a pair of two successive (or query and reference) image frames as input, and then the trained decoder CNN model outputs a warped frame where subtle motions are amplified with an amplification factor. The learned motion representations in the trained encoder-decoder CNN model achieve a better noise-free result than with previous hand-crafted methods. On the other hand, Dorkenwald et al. [28] train the encoder-decoder CNN motion magnification model by unsupervised learning with non-annotated real data for amplifying posture deviations across subjects. This research is variant rather than new because they aim to reveal subtle motion *differences* between two videos rather than subtle motions in a video. While these learning methods have much attention with the recent development of deep learning, they often produce corrupted and/or undesirable results due to strong dependency on their training dataset and also take time to output results; its practical application range is thus still limited.

2.3 Eulerian Video Magnification

2.3.1 Eulerian Video Motion Magnification

Unlike the above Lagrange and learning-based methods, Wu et al. [4] first proposed a video motion magnification method with Eulerian description that measures motions of objects in a video as change in luminance signals over time at a fixed pixel position without object tracking, which is called luminance-based Eulerian video motion magnification (LubEVMM). They show the change in luminance signals represents local motions at the each pixel on the basis of the first-order Taylor series expansions along spatial dimension, which is common in optical flow analysis [19, 29]. This method constructs Laplacian pyramids from image frames to obtain luminance signals at each spatial frequency subband, or each pyramid level. Then, the luminance signals are temporal bandpass filtered with a target temporal frequency and are amplified with an amplification factor to reveal subtle motions at the target temporal frequency in a video. However, this method supports only a small amplification factor at a high spatial frequency subband and often produces noisy results due to directly handling luminance signals in a video.

To overcome these issues, Liu et al. [30] proposed a post-processing method with image warping meshes for refining the LubEVMM results. They first perform LubEVMM and then estimate image-magnifying warping meshes at each time frame by comparing an input video and the magnification video output from LubEVMM. Afterwards, applying the image-magnifying warping meshes to the input video instead of LubEVMM can produce a noise-free motion magnification result because this warping-based post-processing enables us to amplify only subtle motions in a video without touching luminance signals directly. However, this post-processing still produces strange results because it is originally based on the LubEVMM results that contain magnification error.

In contrast with LubEVMM, Wadhwa et al. [5] proposed a current mainstream phase-based Eulerian video motion magnification (PbEVMM) method that measures motions of object in a video as phase signals, which correspond to local motions via the space-shift theorem in the Fourier domain [10], over time frames at a fixed pixel position. This method firstly constructs complex steerable pyramids [31, 32, 33] from image frames to obtain phase signals at each spatial frequency subband and each orientation. Then, the phase signals are temporal bandpass filtered with a target temporal frequency and are amplified with an amplification factor to reveal subtle motions at the target temporal frequency in a video. This PbEVMM method can support a large amplification factor at a high spatial frequency subband and also reduce noise effects in a video compared with the LubEVMM method [4] because the PbEVMM method handles local motions directly but not luminance signals in a video.

The PbEVMM method can produce better magnification results for revealing, e.g., the sway of a bridge, the breathing of an infant, and facial color changes due to blood circulation, than the LubEVMM method. However, the PbEVMM method (or the LubEVMM method) cannot clearly distinguish between subtle phase (or luminance) signals and large ones because their temporal bandpass filtering only focuses on getting the target temporal frequency's components in the phase (or luminance) signals without concerning the existence of large motions in a video. Therefore, the PbEVMM method (and the LubEVMM method) produces messy artifacts when objects in a video move largely.

2.3.2 Eulerian Video Color Magnification

For revealing subtle color changes in a video, Wu et al. [4] first proposed Eulerian video color magnification method (EVCM). This method constructs Gaussian pyramids from image frames to obtain color signals in R, G, B, or Y (this is luminance) color channel at each pyramid level. Then, the color signals at a specific pyramid level (a third pyramid level generally used) are temporal bandpass filtered with a target temporal frequency and are amplified with an amplification factor to reveal subtle color changes at the target temporal frequency in a video. This method produces amazing color magnification results but, as the same of the PbEVMM method [5], also produces messy artifacts when objects in a video move largely due to no concerning large amplitude change in the color signals that often occur at the moment when the background and foreground switch.

2.4 Video Motion Manipulation

As explained in the above sections, the EVMM methods amplify subtle motions in a video to reveal attractive physical/natural phenomena in the small world. However, such manipulating motions to reveal or exaggerate some aspects of a video is not an inherent concept; it has been widely utilized for the computer vision/graphics community.

One of the related interesting work is (de-) animating video. For animating an input video in a simple way, Wand et al. [34] proposed the cartoon animation filter that takes arbitrary motions (such as trajectories with optical flow, motion capture signals, or simple path-based motions created with, e.g., PowerPoint) and modulates them in such a way that the output motions is more alive or animated. The filter is based on a Laplacian of Gaussian filter that enables us to amplify the dominant motions, which has the maximum power spectrum in a video, and/or to generate time-shift motions for producing the important animation effects such as anticipation, follow-through, exaggeration, and squash-and-stretch. In contrast, Bai et al. [35] proposed selectively de-animating video to remove large motions of objects so that other motions are easier to see. While the video motion magnification methods [18, 30, 4, 5] reveal subtle motions by amplifying them, the goal of this method is the opposite because it focuses on revealing the subtle motions by removing larger ones. This method is based on the combination of both the optical flow analysis like the method of [18] to estimate large motions in a video and the graph-cuts [36, 37] to naturally composite video regions, where the original input motions are left or removed, spatially and temporally. This method produces good de-animating video results and facilitates the creation of *cinemagraphs*.

Another one is to modify time-dependent effects of a video captured by high-speed camera. Fuchs et al. [38] use temporal filters to reduce temporal aliasing of motions in a video when the input high-speed video is converted into the low-speed one, and/or to superimpose high-frequency afterimages (caused by fast motions) to the input video so that enables us to understand the motion behaviors without analyzing or tracking. This work can enhance motion display for users to visualize hidden motions captured by high-speed camera.

2.5 Video Motion/Color Analysis

Similar to EVM, analysis of subtle color changes or motions in a video has been researched for various practical applications [39, 40, 41, 42, 1, 2, 3].

In the computer vision community, several works have been proposed for unique practical applications. For example, Davis et al. [39] proposed to recover sounds from subtle vibration of objects, which is caused by the sounds, in high-speed footage. In this work, they explored how the sound-related vibrations vary over an object's surface to confirm what types of the vibration modes of an object can easily recover sounds. Moreover, they proposed to infer material property of objects in a video from the differences of subtle vibration of the objects [40]. This work connects fundamentals of vibration mechanics with the phase-based motion analysis as our explained in Section 2.3.1. Another works [41, 42] analyze motions of objects for estimating geometrical construction. Wadhwa et al. [41] combine the phase-based motion analysis with linear optimization problem for estimating the geometrical distortions of constructions, such as lift bridges and mammalian tectorial membranes. On the other hand, Xue et al. [42] proposed a unique work that reconstructs tree structure from a video. They use physics-based link model with spectrum analysis of twig's vibration because vibration of disconnected branches, though visually similar, often have distinctive natural frequencies. With this model, they can reconstruct tree structure with high accuracy from both tree's spectral vibration and appearance.

Among various practical applications, remote heart rate (HR) detection based on video analysis has been especially researched for medical treatment usage. Verkruysse et al. [1] first proposed a remote method for detecting HR from frontal face videos captured by normal digital camera. This method obtains color signals in a green channel, which strongly represents the absorption of hemoglobin, at each pixel and then averages them over pixels within region of interest for reducing noise effects. After that, HR of a subject is estimated by using the average color signals. This analysis enables us to understand not only HR and also the differences of facial blood flow that can be a symptom of arterial problems. Inspired by [1], Poh et al. [2] extended this method by using blind source separation (BSS) technique, namely independent component analysis, for more robust HR estimation. This approach, which uses the color signals of face for estimating HR, was eagerly improved as described in the survey paper [3] by sophisticating the BSS technique. However, low-rank matrix completion approach [43] has recently adopted to this task that can contribute high accuracy. On the other hand, Balakrishnan et al. [44] proposed another approach for estimating HR based on subtle head motions. They use the subtle head motions, which accompany the cardiac cycle, to extract HR information from a video. This motion analysis approach has a strong advantage that a video view of the head is not restricted and can estimate HR even when skin is not visible. However, since this method has an issue that strongly relies on subtle head motions easily overwhelmed by large ones unrelated to HR estimation, the authors suggest a combination of the subtle motions and color changes will likely prove more useful and robust than using either one independently.

Chapter 3

Preliminary Formulation of Eulerian Video Magnification

In this chapter, we formulate the early EVM methods, the EVCM method [4] as explained in Section 2.3.2 and the PbEVMM method [5] as explained in Section 2.3.1, to clearly explain the following chapters. Note that this formulation is novel in that it is a reinterpretation of the EVM methods by our consideration based on a local Taylor expansion along temporal dimension.

Given a normalized image signal $I(x, y, t) \in [0, 1]$ in one of RGB or YIQ color channels (Y color channel is usually used) at a 2D pixel position (x, y) and a time frame t, they first construct a one-octave Gaussian pyramid $\{I_n(x, y, t) \mid n = 0, \ldots, N-1\}$, which is a set of a color signal $I_n(x, y, t)$ at a pyramid level n for color magnification, or a one/half/quarter-octave (half-octave is usually used) complex steerable pyramid $\{A_{\omega_n,\theta}(x, y, t)e^{i\phi_{\omega_n,\theta}(x,y,t)} \mid n = 0, \ldots, N-1, \theta \in \Theta, 0 \leq \theta < \pi\}$, which is a set of a subband analytic signal $A_{\omega_n,\theta}(x, y, t)e^{i\phi_{\omega_n,\theta}(x,y,t)}$ for motion magnification, where $A_{\omega_n,\theta}(x, y, t)$ is a subband amplitude and $\phi_{\omega_n,\theta}(x, y, t)$ is a subband phase signal at a spatial subband angular frequency ω_n with n and steerable orientation θ in a set of angles Θ.

Here, we define a generalized signal notation $S_{\omega_n,\theta}(x,y,t)$ that represents any signal in an image pyramid with some indexes: a color signal $I_n(x,y,t)$ where $S = I$, $\omega_n := n$, and $\theta = \emptyset$ for color magnification or a subband phase signal $\phi_{\omega_n,\theta}(x,y,t)$ where $S = \phi$ for motion magnification.

Considering a local Taylor expansion along temporal dimension, the early EVM methods [4, 5] assume that $S_{\omega_n,\theta}(x,y,t)$ can be approximated with a first-order local Taylor expansion within a neighborhood of a time $t = h$ as

$$S_{\omega_n,\theta}(x,y,t) \approx S_{\omega_n,\theta}(x,y,h) + \dot{S}_{\omega_n,\theta}(x,y,h)(t-h), \tag{3.1}$$

where $\dot{S}_{\omega_n,\theta}(x,y,h) = \frac{\partial}{\partial t} S_{\omega_n,\theta}(x,y,h)$. With this approximation, the early EVM methods assume that the linear signal $\dot{S}_{\omega_n,\theta}(x,y,h)(t-h)$ represents a subtle color/phase signal, which deviates from the constant signal $S_{\omega_n,\theta}(x,y,h)$, in $S_{\omega_n,\theta}(x,y,t)$. Note that this is why the early EVM methods are often called the Eulerian video linear magnification (EVLM) methods.

Then, to pass a subtle bandpass color/phase signal $C_{\omega_n,\theta,f_t}(x,y,t)$ over time frames at a target temporal frequency f_t, any simple temporal bandpass filter $h(t; f_t)$, such as ideal bandpass filter (IBF) [4, 5] or finite impulse response (FIR) windowed IBF (FIRwinIBF) [5], is convolved with $S_{\omega_n,\theta}(x,y,t)$. Considering Eq. (3.1), let $C_{\omega_n,\theta,f_t}(x,y,t)$ is within a neighborhood of a time $t = h$ and h is the center time of this convolution process, we get

$$C_{\omega_n,\theta,f_t}(x,y,t) = h(t; f_t) * S_{\omega_n,\theta}(x,y,t) \approx \dot{S}_{\omega_n,\theta}(x,y,h)(t-h), \tag{3.2}$$

where $*$ is a convolutional operator, because $h(t; f_t)$ ignores the constant signal $S_{\omega_n,\theta}(x,y,h)$ that indicates the DC component of $S_{\omega_n,\theta}(x,y,t)$ within a neighborhood of a time $t = h$.

After that, $C_{\omega_n,\theta,f_t}(x,y,t)$ multiplied by an amplification factor α is added to $S_{\omega_n,\theta}(x,y,t)$ for obtaining an amplified color/phase signal $\hat{S}_{\omega_n,\theta,f_t}(x,y,t)$ as

$$\hat{S}_{\omega_n,\theta,f_t}(x,y,t) = S_{\omega_n,\theta}(x,y,t) + \alpha C_{\omega_n,\theta,f_t}(x,y,t). \tag{3.3}$$

Finally, we collapse the amplified Gaussian pyramid or the amplified complex steerable pyramid, which is a set of $\hat{S}_{\omega_n,\theta,f_t}(x,y,t)$, to output a magnified image signal $\hat{I}(x,y,t)$ where only subtle color changes or motions at f_t are revealed.

This formulation based on the EVLM methods [4, 5] simply assumes the existence of only subtle color changes or motions in a video as the linear signal $\dot{S}_{\omega_n,\theta}(x,y,h)(t-h)$ in Eq. (3.1). Therefore, the EVLM methods are limited for practical applications as our explained in Chapter 1.

Chapter 4

Ignoring Large Motions in a video

4.1 Introduction

As explained in Section 2.3 and Chapter 3, the early EVM methods, called the EVLM methods [4, 5], use a simple temporal bandpass filter, such as IBF [4, 5] or FIRwinIBF [5], that focuses on only getting the target temporal frequency's components in the phase/color signals. However, as subtle color changes or motions are often obscured by large motions of objects in a real video, the EVLM methods produce messy artifacts when objects in a video move largely because they cannot clearly distinguish between subtle phase/color signals and large ones.

To ignore large motions of objects in a video, layer-based EVLM methods have been proposed [7, 8]. Elgharib et al. [7] require a user to select a region of interest (ROI) where large motions of objects are stabilized by using optical flow that is estimated with translation or affine transformation motion model. Subtle signals left by the stabilization to the ROI are then amplified by the EVLM methods [4, 5]. On the other hand, Kooij et al. [8] automatically select the ROI to be amplified by using a depth-aware bilateral complex steerable pyramid that detects phase signals at each pixel with the same depth value. This depth-aware phase signals can suppress unwanted artifacts caused by large motions between foreground and background. However, these methods require human manipulation [7] or an environment suitable for a depth camera [8]; consequently, these layer-based EVLM methods are time consuming and error prone.

In contrast, Zhang et al. [9] have proposed the EVAM method that attempts to detect subtle color changes or motions in the presence of large motions without the above additional requirements. By assuming that (i) large motions can be approximated linearly in the temporal signal domain and (ii) subtle signals deviate from the linearity, they design temporal acceleration filter (TAF) based on second-order derivative of 1D Gaussian filter. TAF focuses on getting the target temporal frequency's components in the input signals and cuts off linear change in the input signals. This filter can pass only subtle bandpass signals at the target temporal frequency even under the presence of large motions of objects if such motions are slow enough to be linear change in the temporal signal domain. However, this method fails to ignore quick large motions because such motions cause non-linear steep change in the input signals like outlier. Consequently, the EVAM method excessively amplifies quick large motions and produces noisy magnification results in this situation.

In this chapter, we propose a jerk-aware EVAM (JAEVAM) method where our jerk-aware filter (JAF) is applied to the EVAM method [9] for revealing only subtle color changes or motions in the presence of slow and/or quick large motions without the above-mentioned requirements [7, 8]. Our method uses jerk, which has been used to evaluate smoothness of time series data in the neuroscience [11, 12, 13] and mechanical engineering fields [45], to make the EVAM method robust even to quick large motions as well as slow large motions. On the basis of our observation that subtle changes are smoother than quick large motions in the temporal signal domain, we considered that understanding the difference in smoothness enables us to isolate subtle changes from quick large motions. In developing our method, we used jerk-based smoothness to design JAF that only passes subtle changes in the presence of quick large motions. By applying our JAF to the EVAM method, we obtain impressive magnification results without messy artifacts, which could be caused by slow and/or quick large motions of objects, in both real and synthetic videos.

4.2 Conventional Method: Eulerian Video Acceleration Magnification [9]

Unlike the EVLM methods described in Chapter 3, the Eulerian video acceleration magnification (EVAM) method [9] assumes that $S_{\omega_n,\theta}(x,y,t)$ can be approximated with a second-order local Taylor expansion within a neighborhood of a time $t = h$ as

$$\begin{aligned} S_{\omega_n,\theta}(x,y,t) \approx S_{\omega_n,\theta}(x,y,h) + \dot{S}_{\omega_n,\theta}(x,y,h)(t-h) \\ + \frac{1}{2}\ddot{S}_{\omega_n,\theta}(x,y,h)(t-h)^2, \end{aligned} \tag{4.1}$$

where $\ddot{S}_{\omega_n,\theta}(x,y,h) = \frac{\partial^2}{\partial t^2} S_{\omega_n,\theta}(x,y,h)$. With this approximation, the EVAM method assumes that the linear signal $\dot{S}_{\omega_n,\theta}(x,y,h)(t-h)$ represents large motions of objects in spatial domain and the non-linear quadratic signal $\frac{1}{2}\ddot{S}_{\omega_n,\theta}(x,y,h)(t-h)^2$ represents a subtle color/phase signal in $S_{\omega_n,\theta}(x,y,t)$.

Considering Eq. (4.1), the EVAM method proposed a temporal acceleration filter (TAF) as a temporal bandpass filter. TAF is a combination of the second-order derivative operator $\frac{\partial^2}{\partial t^2}$ and the Gaussian function $G_\sigma(t)$ with its standard deviation σ, defined as

$$\text{TAF}(t; f_t) = \frac{\partial^2 G_\sigma(t)}{\partial t^2}. \tag{4.2}$$

The σ determines f_t based on scale-space theory [46, 47] and is set as $\sigma = \frac{f_s}{4f_t\sqrt{2}}$

with the video sampling frame rate f_s. This filter can get the target temporal frequency's components in the input signals by controlling σ and completely cuts off the constant signal $S_{\omega_n,\theta}(x,y,h)$ and the linear signal $\dot{S}_{\omega_n,\theta}(x,y,h)(t-h)$ because of $\frac{\partial^2}{\partial t^2}$ operator.

Consequently, to pass a subtle bandpass color/phase signal $C_{\omega,\theta,f_t}(x,y,t)$ over time frames at a target temporal frequency f_t, the EVAM method convolves TAF$(t;f_t)$ to $S_{\omega_n,\theta}(x,y,t)$. Here, considering Eq. (4.1), let $C_{\omega_n,\theta,f_t}(x,y,t)$ is within a neighborhood of a time $t=h$ and h is the center time of this convolution process, we get

$$C_{\omega_n,\theta,f_t}(x,y,t) = \text{TAF}(t;f_t) * S_{\omega_n,\theta}(x,y,t) \approx \frac{1}{2}\ddot{S}_{\omega_n,\theta}(x,y,h)(t-h)^2. \qquad (4.3)$$

After that, the EVAM method follows the same process of the EVLM methods [4, 5] as described in Eq. (3.3) and its below.

This EVAM method produces good magnification results even under the presence of large motions of objects if such motions are slow enough to be linear change in the input signal, described as $\dot{S}_{\omega_n,\theta}(x,y,h)(t-h)$, because this linear signal can be completely cut off by $\frac{\partial^2}{\partial t^2}$ operator in TAF. However, this method cannot ignore quick large motions because such motions cause non-linear steep change in the input signal like outlier, which is unfortunately included in the non-linear quadratic signal $\frac{1}{2}\ddot{S}_{\omega_n,\theta}(x,y,h)(t-h)^2$ that is assumed to be a subtle color/phase signal in $S_{\omega_n,\theta}(x,y,t)$. Therefore, this method excessively amplifies quick large motions and produces noisy magnification results in this practical situation.

4.3 Proposed Method: Jerk-Aware Eulerian Video Acceleration Magnification [48]

To reveal only subtle color changes or motions in the presence of slow and/or quick large motions, we propose a jerk-aware EVAM (JAEVAM) method that applies our jerk-aware filter (JAF) to EVAM method [9] to make it robust even to quick large motions as well as slow large motions. First, we argue that jerk is a useful index to handle quick large motions in EVM. Second, we describe how we designed JAF that passes subtle color changes or motions only under quick large motions by using jerk-based smoothness. Finally, we show how we applied this filter to the EVAM method.

4.3.1 Jerk

As our mentioned before, the EVAM method [9] assumes that slow large motions are approximately linear in the temporal signal domain, whereas our key idea is based on our observation that subtle color changes or motions depict smoother trajectories than quick large motions in the temporal signal domain (Fig. 4.1). We argue that understanding the difference in the signal smoothness better enables us to isolate subtle color changes or motions from quick large motions. Therefore, we focus on a differential feature called jerk.

Jerk is a third temporal derivative of displacement, and represents the rate of change in acceleration per unit of time. It is an effective index to assess steepness or smoothness of time series data; its value becomes high during steep changes but low during smooth changes. It has been used in many research fields for assessing movements and trajectories [11, 12, 13, 45, 49]. In neuroscience, it has been used to model the trajectory of voluntary arm movements [11] or to assess the recovery of motor performance in stroke patients [12, 13]. In mechanical fields, the trajec-

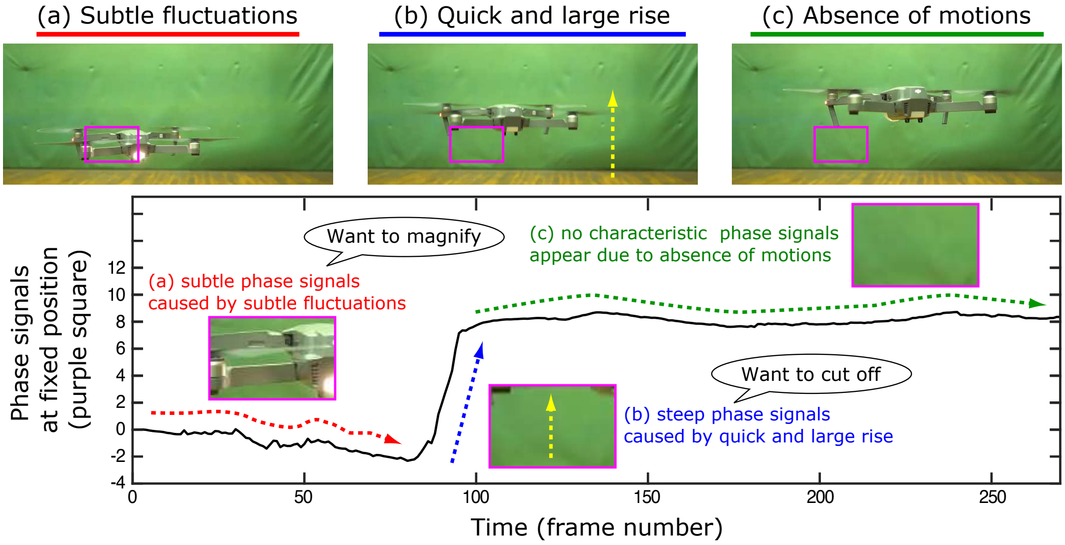
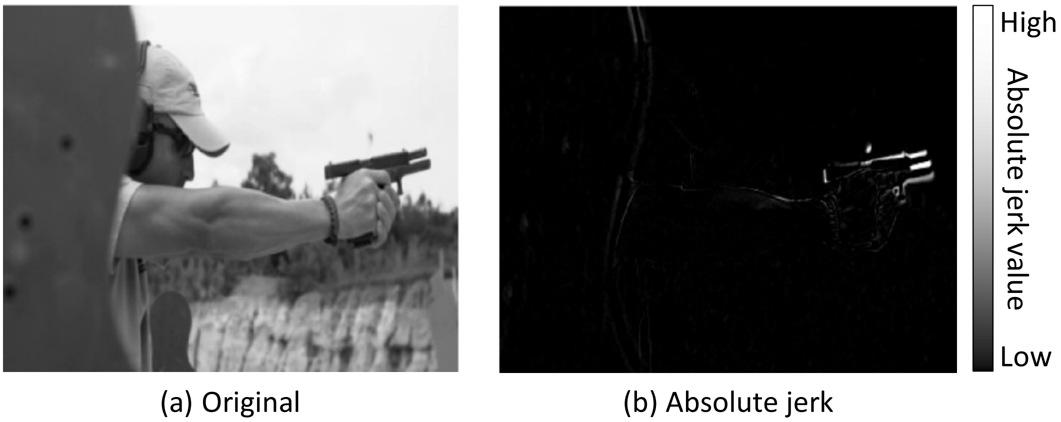


Figure 4.1: Our observation. The EVM methods measure color changes or motions in a video as color/phase signals at a fixed position (purple squares in (a), (b), and (c)). We observed that subtle phase signals caused by subtle fluctuations of the drone (a) are smoother than non-linear steep changes in the phase signals caused by quick and large rise motions of the drone (b) because its magnitude is very small.

tories of robot models with jerk restrictions make it possible to obtain smooth control [45]. Through these findings, we assume that subtle color changes or motions in the temporal signal domain have a lower jerk value than quick large motions due to difference of the smoothness. To verify our hypothesis, we simply checked the third temporal derivative of the luminance signals in a gun-shooting video (Fig. 4.2). As a result, static objects (e.g. body and arm) having imperceptible smooth subtle deformations or slow smooth camera panning (background) show lower jerk values than those in quick motions of objects (e.g. gun and cartridge).



(a) Original

(b) Absolute jerk

Figure 4.2: Gun-shooting video in luminance space (a) and jerk calculated by luminance signals (b). Jerk only responds to quick large motions, such as gun blowback and gun cartridge release. On the other hand, it does not respond to static objects (e.g. body, arm, and background) that seem to have imperceptible smooth subtle deformations (body and arm) or slow smooth camera panning (background).

4.3.2 Jerk-Aware Filter (JAF)

To pass only subtle color/phase signals in the presence of quick large motions, we attempted to design JAF from our knowledge of jerk. This filter was designed to have jerk-based smoothness so that it will pass subtle color changes or motions only and cut off quick large motions.

Given an input color/phase signal $S_{\omega_n,\theta}(x,y,t)$, we first calculate a bandpass jerk signal $J_{\omega_n,\theta,f_t}(x,y,t)$ with f_t by convolving a third-order derivative of the Gaussian function as

$$J_{\omega_n,\theta,f_t}(x,y,t) = \frac{\partial^3 G_\sigma(t)}{\partial t^3} * S_{\omega_n,\theta}(x,y,t), \tag{4.4}$$

where σ is set as that it in Eq. (4.2) of the EVAM method [9].

Second, we transform $J_{\omega_n,\theta,f_t}(x,y,t)$ into a bandpass jerk-based smoothness $\hat{J}_{\omega_n,\theta,f_t}(x,y,t)$ that has a high value (close to 1) when a smooth signal appears and a low value (close to 0) when no such a signal appears as

$$\begin{aligned}\hat{J}_{\omega_n,\theta,f_t}(x,y,t) &= 1 - \Gamma\left(J_{\omega_n,\theta,f_t}(x,y,t)\right), \\ \Gamma\left(J_{\omega_n,\theta,f_t}(x,y,t)\right) &= \frac{|J_{\omega_n,\theta,f_t}(x,y,t)| - \min\limits_{x,y,t}(|J_{\omega_n,\theta,f_t}(x,y,t)|)}{\max\limits_{x,y,t}(|J_{\omega_n,\theta,f_t}(x,y,t)|) - \min\limits_{x,y,t}(|J_{\omega_n,\theta,f_t}(x,y,t)|)},\end{aligned} \tag{4.5}$$

where $\Gamma(\cdot)$ is the min-max normalization function with respect to x, y, and t.

Finally, to provide a filter capable of easily adjusting the effect of the jerk-based smoothness, we add an exponent parameter $\beta > 0$ to Eq. (4.5) and then design our JAF as $\text{JAF}_{\omega_n,\theta}(x,y,t;\sigma,\beta)$, which can selectively pass subtle color/phase signals in the presence of quick large motions:

$$\text{JAF}_{\omega_n,\theta}(x,y,t;f_t,\beta) := \hat{J}_{\omega_n,\theta,f_t}(x,y,t)^{\beta}. \tag{4.6}$$

4.3.3 Pyramid-based Correction

The EVM methods construct the complex steerable (or Gaussian) pyramid to perform their algorithm at each pyramid level n [4, 5, 7, 8, 9]. Under the pyramid construction process, we should re-consider the meaning of $\text{JAF}_{\omega_n,\theta}(x,y,t;f_t,\beta)$ in terms of a pyramid level n. As mentioned in previous studies [17, 50], image sequences at higher n can handle large displacements, but their values decrease in proportion to the image resolution at n. This means that though JAF at higher n capture much quicker large motions, it is underestimated due to the low resolution at higher n. Therefore, we define a filter correction based on the pyramid scaling factor λ_n, where λ_n represents a scaling down ratio of the image resolution at each n against the original image resolution, as

$$\text{JAF}_{\omega_n,\theta}(x,y,t;f_t,\beta,\lambda_n) := \text{JAF}_{\omega_n,\theta}(x,y,t;f_t,\beta)^{1/\lambda_n}. \tag{4.7}$$

Furthermore, we consider that JAF will need to be modified by using a similar coarse-to-fine approach [17, 50]. As image sequences at higher n detect image changes in a wider space, they can accurately capture quick large motions and calculate correct jerk. However, quick large motions do not fit in the detection space at lower n. This means that jerk at lower n cannot reflect this essence correctly even if quick large motions occur. Therefore, it is necessary to propagate the information of JAF at higher n to that at lower one. We define this propagation correction as

$$\mathrm{JAF}_{\omega_n,\theta}(x, y, t; f_t, \beta, \lambda_n, V) := \prod_{i=n}^{n+V} \mathcal{R}\left(\mathrm{JAF}_{\omega_i,\theta}(x, y, t; f_t, \beta, \lambda_i), n\right), \tag{4.8}$$

where V is the number of the pyramid level for this correction, and the function of $\mathcal{R}(\cdot, n)$ resizes the size of JAF to that at n with bicubic interpolation. Through this correction, our JAF can effectively pass only subtle color/phase signals in the presence of quick large motions.

4.3.4 Applying JAF to the EVAM method

We additionally apply $\mathrm{JAF}_{\omega_n,\theta}(x, y, t; f_t, \beta, \lambda_n)$ of Eq. (4.7) to Eq. (4.3) for color magnification, and $\mathrm{JAF}_{\omega_n,\theta}(x, y, t; f_t, \beta, \lambda_n, V)$ of Eq. (4.8) to Eq. (4.3) for motion magnification. Through applying JAF to the EVAM method, we obtain a result where only subtle color changes or motions are revealed under the presence of slow and quick large motions.

4.4 Theoretical View for Proposed Method

Unlike the EVLM [4, 5] methods and the EVAM [9] method, we argue that $S_{\omega_n,\theta}(x, y, t)$ can be *strictly* decomposed by a second-order local Taylor expansion within a neighborhood of a time $t = h$ with an approximation error as

$$\begin{aligned} S_{\omega_n,\theta}(x, y, t) = S_{\omega_n,\theta}(x, y, h) &+ \dot{S}_{\omega_n,\theta}(x, y, h)(t - h) \\ &+ \frac{1}{2}\ddot{S}_{\omega_n,\theta}(x, y, h)(t - h)^2 + R_{2,\omega_n,\theta}(x, y, t), \end{aligned} \tag{4.9}$$

where $R_{2,\omega_n,\theta}(x,y,t)$ indicates the approximation error of the second-order local Taylor expansion of Eq. (4.1) and can be strict defined via the well-known Lagrange form as

Theorem 1. *If $S(t)$ is continuous on a closed interval between t and h, there exists a number c between t and h such that*

$$R_k(t) = \frac{1}{(k+1)!}\frac{\partial^{(k+1)}S(c)}{\partial t^{(k+1)}}(t-h)^{k+1}. \tag{4.10}$$

Therefore,

$$R_{2,\omega_n,\theta}(x,y,t) = \frac{1}{6}\dddot{S}_{\omega_n,\theta}(x,y,c)(t-h)^3, \tag{4.11}$$

where $\dddot{S}_{\omega_n,\theta}(x,y,c) = \frac{\partial^3}{\partial t^3}S_{\omega_n,\theta}(x,y,c)$. By comparing Eq. (4.1) and Eq. (4.9), Eq. (4.1) assumes all large motions are slow enough to be approximately the linear signal $\dot{S}_{\omega_n,\theta}(x,y,h)(t-h)$, but Eq. (4.9) assumes there are large motions that can be a non-linear steep signal $R_{2,\omega_n,\theta}(x,y,t)$. Therefore, we can assume that the non-linear steep signal $R_{2,\omega_n,\theta}(x,y,t)$, rather than the non-linear quadratic signal $\frac{1}{2}\ddot{S}_{\omega_n,\theta}(x,y,h)(t-h)^2$, represents quick large motions of objects in spatial domain.

Considering this strict decomposition of Eq. (4.9) and Eq. (4.11), let the band-pass jerk signal $J_{\omega_n,\theta,f_t}(x,y,t)$ of Eq. (4.4) is within a neighborhood of a time $t = h$ and h is the center time of the convolution process of Eq. (4.4), we get

$$\begin{aligned} J_{\omega_n,\theta,f_t}(x,y,t) &= \frac{\partial^3 G_\sigma(t)}{\partial t^3} * S_{\omega_n,\theta}(x,y,t) \\ &= G_\sigma(t) * \dddot{S}_{\omega_n,\theta}(x,y,t) \\ &\approx \frac{1}{6}\dddot{S}_{\omega_n,\theta}(x,y,c)(t-h)^3 \\ &= R_{2,\omega_n,\theta}(x,y,t). \end{aligned} \tag{4.12}$$

Therefore, our JAF of Eq. (4.6) can be approximated with $R_{2,\omega_n,\theta}(x,y,t)$ as

$$\text{JAF}_{\omega_n,\theta}(x,y,t;f_t,\beta) \approx \left(1 - \Gamma\left(R_{2,\omega_n,\theta}(x,y,t)\right)\right)^\beta. \tag{4.13}$$

This approximation indicates that $\text{JAF}_{\omega_n,\theta}(x, y, t; f_t, \beta)$ can be interpreted as the inverse criteria of the approximation error $R_{2,\omega_n,\theta}(x, y, t)$ that the EVAM method unfortunately has.

From Eq. (4.13), through applying $\text{JAF}_{\omega_n,\theta}(x, y, t; f_t, \beta)$ to the convolution process of Eq. (4.3), we have

$$\begin{aligned}&\text{JAF}_{\omega_n,\theta}(x, y, t; f_t, \beta) \cdot \big(\text{TAF}(t; f_t) * S_{\omega_n,\theta}(x, y, t)\big)\\ &\approx \begin{cases} C_{\omega_n,\theta,f_t}(x, y, t), & \Gamma\left(R_{2,\omega_n,\theta}(x, y, t)\right) \approx 1, \\ 0, & \Gamma\left(R_{2,\omega_n,\theta}(x, y, t)\right) \approx 0. \end{cases}\end{aligned} \tag{4.14}$$

This equation indicates that JAF with TAF never passes all signals when $\Gamma\left(R_{2,\omega_n,\theta}(x, y, t)\right)$ is close to 1, which means non-linear steep signals caused by quick large motions are dominant in the input signal $S_{\omega_n,\theta}(x, y, t)$, or passes only subtle color/phase signals $C_{\omega_n,\theta,f_t}(x, y, t)$ when $\Gamma\left(R_{2,\omega_n,\theta}(x, y, t)\right)$ is close to 0, which means there is no quick large motions in $S_{\omega_n,\theta}(x, y, t)$.

4.5 Experiments and Results

4.5.1 Experimental Setup

To evaluate the effectiveness of our proposed method, we conducted experiments on real videos as well as on synthetic ones with ground truth magnification. We assessed the effectiveness qualitatively with the real videos and it quantitatively with synthetic videos against the ground truth. We set the amplification factor α, the target temporal frequency f_t to be amplified, and hyper parameter β of Eq. (4.6) as given in Table 4.1. We applied our proposed method and comparison methods to a video in the YIQ color space.

Table 4.1: Experimental parameters for all videos. We selected the all parameters to be the best ones for each experiment.

Video	f_t	fs	β	α (ours,[4],[5],[9],[27])
Light bulb 1 [9]	10	600	0.0001	(25, 25, –, 25, –)
Light bulb 2 [48]	2	160	20	(40, 40, –, 40, –)
Golf [48, 51]	2	60	0.8	(20, –, 12, 12, 12)
Gun [9]	20	480	0.3	(10, –, 8, 8, 6)
Drone [48, 51]	2	30	1	(25, –, 18, 18, 6)
Ukulele [48, 51]	40	240	1	(25, –, 18, 18, 12)
Synthetic ball	10	60	0.0001-5	(35, –, 20, 20, 4)

Color Magnification. We constructed a Gaussian pyramid to decompose each image frame and amplified Y color signals only on the third level of the pyramid. This approach is similar to that used in [4, 9].

Motion Magnification. To decompose each image frame into each subband analytic signal, we constructed a complex steerable pyramid [5] with half-octave bandwidth filters and eight orientations in Y color space. We empirically set the number of propagation V as 5 in propagation correction of Eq. (4.8), and this correction was done independently for each orientation.

4.5.2 Color Magnification in Real Videos

Figure 4.3 shows a light bulb slowly moving upward. The color changes in the light bulb caused by the electrical current changing are hardly visible without magnification (see the original in Fig. 4.3). The EVLM method [4] produces clipping artifacts due to the slow large motions. In contrast, the EVAM method [9] and our method clearly magnify subtle color changes in the light bulb under the slow large

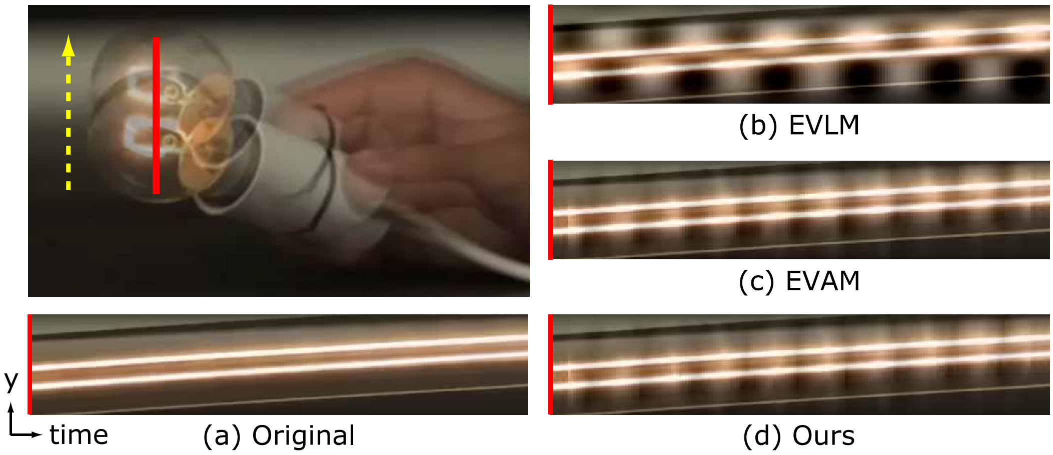


Figure 4.3: Color magnification in the presence of slow large motions; light bulb slowly moves upward. (a) Original video. (b) The EVLM method [4]. (c) The EVAM method [9]. (d) Our method. The EVAM method [9] and our method effectively magnify subtle color changes in the light bulb under the slow large motions.

motions. These results suggest that our method did not have any negative effects on the video that the EVAM method [9] produced good color magnification results in the presence of slow large motions.

Figure 4.4 shows light bulbs shattered by a bullet shot from a gun. Processing this video with the EVLM method [4] reveals color changes but creates clipping artifacts. While the EVAM method [9] succeeds in clearly magnifying subtle color changes, it detects steep color changes due to quick-flying transparent fragments of the broken light bulbs and produces messy artifacts. In contrast, our method can magnify only subtle color changes before and after the light bulbs shattering despite the quick-flying transparent fragments.

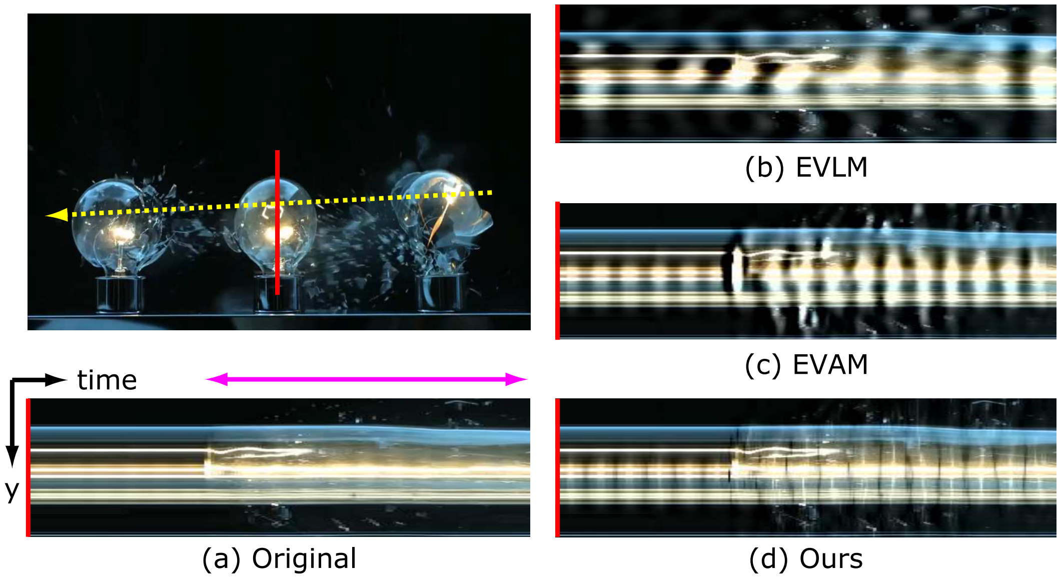


Figure 4.4: Color magnification in the presence of quick large motions; light bulbs shattered by a gun bullet which is depicted by the yellow arrow. Our method only magnifies subtle color changes in the light bulbs during the quick-flying transparent fragments of the broken light bulbs (see the purple arrow time intervals).

4.5.3 Motion Magnification in Real Videos

Figure 4.5 shows the motion magnification results for a golf swing video to magnify the subtle deformation of the iron shaft that occurs when the ball is hit. The EVLM method [5] induces large artifacts due to the quick large swing motion. The EVAM method [9] can magnify the subtle deformation of the iron shaft that occurs when the ball is hit, but it induces collapsing of the shape of the iron shaft due to the quick large swing motion. The LebVMM method [27] can magnify the subtle deformation of the iron shaft but constantly induces strange magnified deformation. Our method can reveal this deformation by magnifying the subtle deformation of the iron shaft and ignoring the effects of the quick large swing motion.

Figure 4.6 shows a gun-shooting video with slow camera panning and quick gun

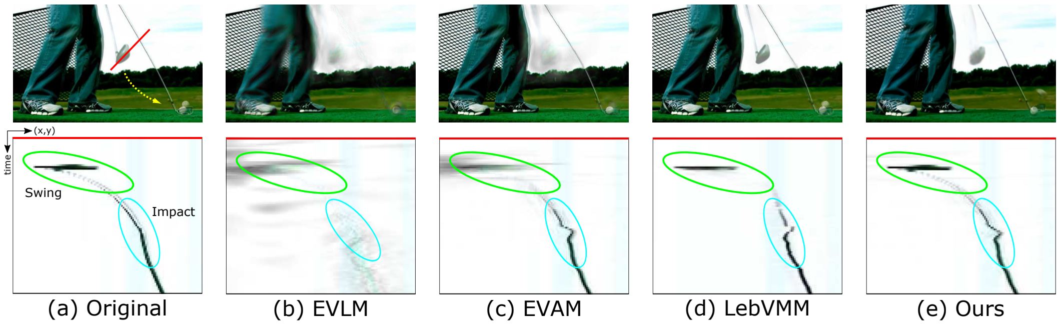


Figure 4.5: Visualization of impact spread in the iron shaft of golf club. In the left top, the yellow arrow depicts the golf swing along a trajectory. The top row shows 2 frames overlaid to indicate the swing phase and the impact phase of the ball. The bottom row shows the spatiotemporal slices along a single diagonal red line on the top of row of (a); the green circles indicate the swing phase, and the cyan circles indicate the impact phase. (a) Original video. (b) The EVLM method [5]. (c) The EVAM method [9]. (d) The LebVMM method [27]. (e) Our method. Our method only magnifies subtle deformation of the iron shaft without artifacts caused by quick swinging motions as with other methods. See the supplementary material for the video results.

recoil motion. We magnify the subtle deformation of the muscles and the skin due to the strong gun recoil. The EVLM method [5] induces large noise due to the slow camera panning and quick gun recoil motion. The EVAM method [9] can magnify subtle skin deformation of the arm in the presence of slow camera panning but induces collapse of the gun shape due to misdetected quick gun recoil motion. The LebVMM method [27] can magnify subtle skin deformation of the arm but induces disappearance of the gun tip due to quick gun recoil motions. Our method magnifies only the skin deformation of the arm in the presence of slow camera panning and quick gun recoil motion.

Figure 4.7 shows an example of applying our proposed magnification to reveal

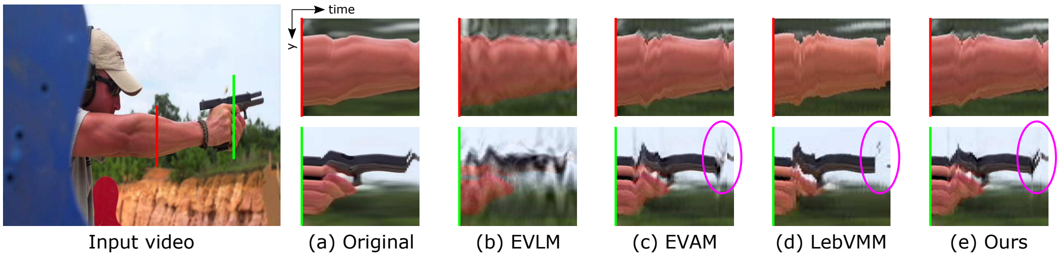


Figure 4.6: Visualization of impact spread throughout an athlete's body. Spatiotemporal slices are shown along a single red and green lines (top-left). Our method magnifies subtle deformations in the arm without the effect of camera panning or gun recoil motion (see the purple circles).

autonomous fluctuations of the drone during flight. In this case, a drone subtly fluctuates with various large motions: slow parallel transition, quick rising, and 3D rotation of the body shift. The EVLM method [5] produces large artifacts due to the magnification of all the drone's motions. The EVAM method [9] can magnify the subtle fluctuations of the drone under the slow parallel transition but induces shape collapses of the drone due to the quick rising and strong quick light flickering, which is misdetected as quick large motions in the texture. Similar to the EVAM method [9], the LebVMM method [27] can magnify the subtle fluctuations of the drone but induces shape collapses of the drone due to the quick rising. Our method magnifies only the subtle fluctuations of the drone without the effects of the various large motions.

Figure 4.8 shows the case for a ukulele strumming video in which quick hand motions occur several times. The EVLM method [5], the EVAM [9], and the LebVMM method [27] produce artifacts around the quick strumming hand motions, but our method automatically ignores all the quick strumming hand motions and can magnify the subtle vibrations of the ukulele strings without user annotations or additional information.

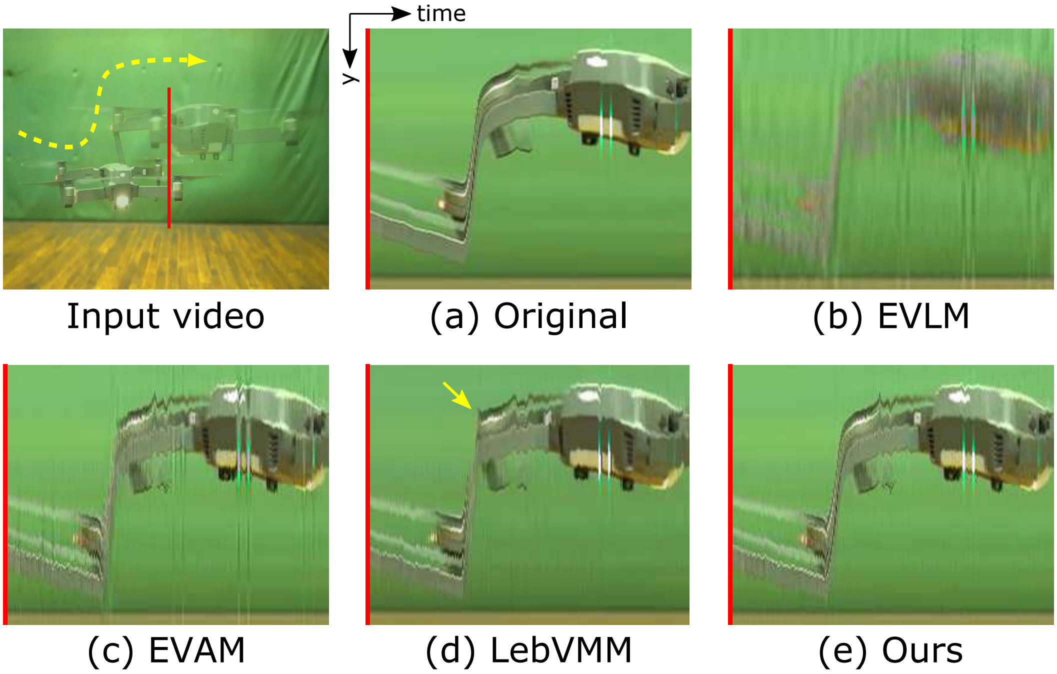


Figure 4.7: Visualization of autonomous fluctuations of the drone during flight. Spatiotemporal slices are shown along a single vertical red lines (left). Our method magnifies and reveals the subtle fluctuations of the drone without artifacts caused by various large motions.

4.5.4 Controlled Experiments

In Figure 4.9 (left), we show a 4-second synthetic ball video. We set the radius of the ball as 20 pixels. The ball has vertically subtle motions defined as $d_{\text{subtle}} = A_{\text{subtle}} \sin\left(2\pi \frac{f_t}{f_s} j\right)$ where $A_{\text{subtle}} = 0.5$ pixels, $f_t = 10$ cycles/frame, $f_s = 60$ frames/second, and j is the frame number. The ball also has vertically slow large motions on the screen from the top to the bottom, with $d_{\text{slow}} = 0.5$ pixels/frame. When the frame number j reaches 80 frames, the ball moves quickly and horizontally with $d_{\text{quick}} = A_{\text{quick}} \sin\left(2\pi \frac{f_{\text{quick}}}{f_s} j\right)$ where $A_{\text{quick}} = [0, 100]$ pixels, $f_{\text{quick}} = 2$ cycles/frame, but after 20 frames, returns to how it was before. To obtain the ground truth of the subtle motion magnification, we created a true magnification video while changing d_{subtle} to $d_{\text{subtleMag}} = 5 \cdot d_{\text{subtle}}$.

Firstly, we assessed the effectiveness of each motion magnification method for magnifying the subtle motions and ignoring the quick horizontal large motions of the synthetic ball while changing A_{quick} relative to the ground truth video. We applied the different EVM methods to this video. We fixed $\beta = 1$ for all methods that required it, and each α as listed in Table 4.1.

Note that to investigate the effectiveness of our proposed method using $\text{JAF}_{\omega_n,\theta}(x, y, t; f_t, \beta, \lambda_n, V)$ of Eq. (4.8) in terms of pyramid-based correction as explained in Subsection 4.3.3, we prepared two methods: a jerk method that uses $\text{JAF}_{\omega_n,\theta}(x, y, t; f_t, \beta)$ of Eq. (4.6) and a jerk-down method that uses $\text{JAF}_{\omega_n,\theta}(x, y, t; f_t, \beta, \lambda_n)$ of Eq. (4.7).

Figure 4.9 (right) shows the mean square error (MSE) we obtained between each magnification result and the ground truth motion magnification as $A_{\text{quick}} = 100$ measured in each frame. For the EVLM method [5], we magnified the vibration in the frequency range of 9 to 11 Hz. This method incurs major errors in all the frames due to slow and quick large motions. The EVAM method [9] and the LebVMM method [27] magnify subtle motions when slow large motions appear but produces extremely major errors when quick large motions appear. The jerk and jerk-down methods can cope with quick horizontal large motions fairly well, but our proposed method, despite its bigger amplification factor, best handles quick horizontal large motions while magnifying subtle motions that resemble the ground truth in all the frames.

Figure 4.10 shows how a synthetic ball video behaves with different quick horizontal large motions A_{quick}. At each A_{quick}, we calculated the mean MSE when subtle motions appear with slow large motions (mMSE$_{\text{subtle}}$) and the mean MSE when quick horizontal large motions appear (mMSE$_{\text{quick}}$) relative to the ground truth. The

EVLM method [5] has large mMSE$_{\text{subtle}}$ and mMSE$_{\text{quick}}$ for every A_{quick} due to the effects of slow and quick large motions. The EVAM method [9] keeps mMSE$_{\text{subtle}}$ low, but mMSE$_{\text{quick}}$ increases in proportion to A_{quick}. The LebVMM method [27] behaves the same as method [9] but shows slightly higher mMSE$_{\text{subtle}}$ and mMSE$_{\text{quick}}$ due to the strong dependence of the training dataset. The jerk method performs better than the above state-of-the-art methods but mMSE$_{\text{quick}}$ still increases in proportion to A_{quick}. The jerk-down method keeps mMSE$_{\text{quick}}$ lower than the above four methods, but our proposed method is the best at keeping mMSE$_{\text{subtle}}$ and mMSE$_{\text{quick}}$ low even when A_{quick} is increasing.

Secondly, to evaluate the effectiveness of our proposed method in terms of pyramid-based correction, we applied the jerk method, the jerk-down method, and our proposed method to the synthetic ball video with $A_{\text{quick}} = 100$ while changing β to 0.0001 and 5

Figure 4.11 shows mMSE$_{\text{subtle}}$ and mMSE$_{\text{quick}}$ for every β relative to the ground truth video. Although β increased in this case, the jerk method was not able to handle the quick large motions well (Fig. 4.11 left). As we added down sampling correction to our JAF, the jerk-down method correctly obtained the value of quick large motions in proportion to l. Thus, this method can obtain lower mMSE$_{\text{subtle}}$ and mMSE$_{\text{quick}}$. However, it cannot completely ignore quick large motions; as the center of Figure 4.11 shows, mMSE$_{\text{quick}}$ does not reach 0. Our proposed method uses our JAF with all pyramid corrections: down sampling correction and propagation correction. By integrating spatial information across the pyramid hierarchy through propagation correction, our method puts mMSE$_{\text{quick}}$ at almost 0 and keeps mMSE$_{\text{subtle}}$ low; this implies our proposed method best handles quick large motions and magnifies subtle motions in the presence of slow large motions without user annotations or additional information (Fig. 4.11 right).

Finally, we evaluated the validity of our JAF in handling quick large motions. As our mentioned before in Section 4.4, our JAF can represent the approximation error $R_{2,\omega_n,\theta}(x, y, t)$ of the EVAM method [9]. However, we considered that $C_{\omega_n,\theta,f_t}(x, y, t)$ obtained from Eq. (4.3) can be also used to represent $R_{2,\omega_n,\theta}(x, y, t)$ although it includes both the target non-linear quadratic signal $\frac{1}{2}\ddot{S}_{\omega_n,\theta}(x, y, h)(t-h)^2$ and $R_{2,\omega_n,\theta}(x, y, t)$. We were convinced that our JAF is better than a filter designed by $C_{\omega_n,\theta,f_t}(x, y, t)$ because ours is designed with considering only $R_{2,\omega_n,\theta}(x, y, t)$. However, to precisely evaluate the effectiveness of our JAF, we designed an acceleration-aware filter by converting $J_{\omega_n,\theta,f_t}(x, y, t)$ in Eq. (4.5) to $C_{\omega_n,\theta,f_t}(x, y, t)$ obtained from Eq. (4.3), which has a low value (close to 0) when $C_{\omega_n,\theta,f_t}(x, y, t)$ is high and a high value (close to 1) when $C_{\omega_n,\theta,f_t}(x, y, t)$ is close to 0.

To compare our JAF and the acceleration-aware filter, we prepared 1D signals $\sin\left(2\pi\frac{1}{1000}j\right) + 0.1\sin\left(2\pi\frac{20}{1000}j\right)$ in which the first term indicated linear signals caused by slow large motions and the second term indicated subtle signals. When the frame number j reached the 700 frame, we added steep signals caused by quick large motions. We defined "subtle time" as when subtle signals appeared and "steep time" as when steep signals appeared. To obtain the ground-truth magnification for the 1D signals, we created another 1D signals as $\sin\left(2\pi\frac{1}{1000}j\right) + 0.5\sin\left(2\pi\frac{20}{1000}j\right)$, where the second term were amplified 5 times.

The top-left panel in Figure 4.12 shows the original signals (black) and the ground-truth magnification signals (red). The other panels show the original signals (black) and the magnification result (green, blue or purple) produced by each magnification method. Note that the each filter weight was set to $\beta = 60$. The EVAM method [9] can magnify the subtle signals but excessively magnify the steep signals at the 700 frame (Fig. 4.12, top-right). Applying the acceleration-aware filter to EVAM method can suppress the steep signals but also accidentally suppress the subtle signals (Fig. 4.12, bottom-left). On the other hand, our proposed method, in which our JAF is combined with method, can suppress only the steep signals and amplify the subtle signals (Fig. 4.12, bottom-right).

In Figure 4.13, we calculated the mean MSE during the subtle time (mMSE$_{\text{subtle}}$) and during the steep time (mMSE$_{\text{steep}}$) at each $\alpha = [0, 200]$ and $\beta = [0, 100]$, relative to the ground truth. The EVAM method [9] with the acceleration-aware filter needs to search for α and β simultaneously to obtain the lowest mMSEs in both the subtle and the steep times because the acceleration-aware filter includes the target subtle signals (Fig. 4.13, top). On the other hand, our method of using the EVAM method [9] with JAF only searches for α and β independently to obtain the lowest mMSEs in both the subtle and the steep times; the two parameters are almost independent in the parameter space where mMSEs are low (Fig. 4.13, bottom). These results indicate the acceleration-aware filter negatively affects the amplified subtle signals in method while JAF does not; JAF focuses on suppressing only the steep signals caused by quick large motions. Therefore, our proposed method purely extends method [9] to deal with the quick large motions of objects.

4.6 Discussions and Limitations

While our proposed method expands the applicable range of video magnification by overcoming the disturbance of quick large motions, it has limitations.

Our JAF can cut off quick large motions while permeating subtle color changes or motions. However, if subtle color changes or motions with quick large motions appear, our method weakly magnify or ignore the subtle changes. For example, Figure 4.5 shows that our method can magnify subtle deformations of the iron shaft that occur when the ball is hit, but cannot magnify them while the club is being swung. This is due to the fact that the subtle deformations mixed with quick large swing motions are subject to removal by JAF. However, such motions are out of the range of our magnified targets. Even if we can magnify such subtle motions, quick large motions overwhelm these magnification results and we cannot follow them with the naked eye. Developing a method for detecting and magnifying subtle color changes or motions mixed with quick large motions can be a subject for future work.

Another limitation of our method is that it makes subtle color changes or motions slightly sharp. Figure 4.4 shows that our method slightly sharpens the subtle color changes and shortens the time intervals of the light changes compared with the EVAM method [9]. In Figure 4.12, our method amplifies only the subtle signals but slightly distorts the shape of the smooth signals as a sawtooth shape. This effect appears as an increase in the MSE in Figures 4.9 to 4.11; our method slightly increases MSE when subtle changes appear. However, this is not a serious problem for video magnification because the most important point is to detect and magnify the amplitude of subtle signals to reveal subtle color changes or motions in a video. As shown in Figures 4.12 and 4.13, our method is superior in that it can filter out only the steep signals caused by quick large motions without affecting the amplitude of the subtle signals. However, an unresolved problem, which is for future work, is that our proposed method slightly distorts magnified subtle changes.

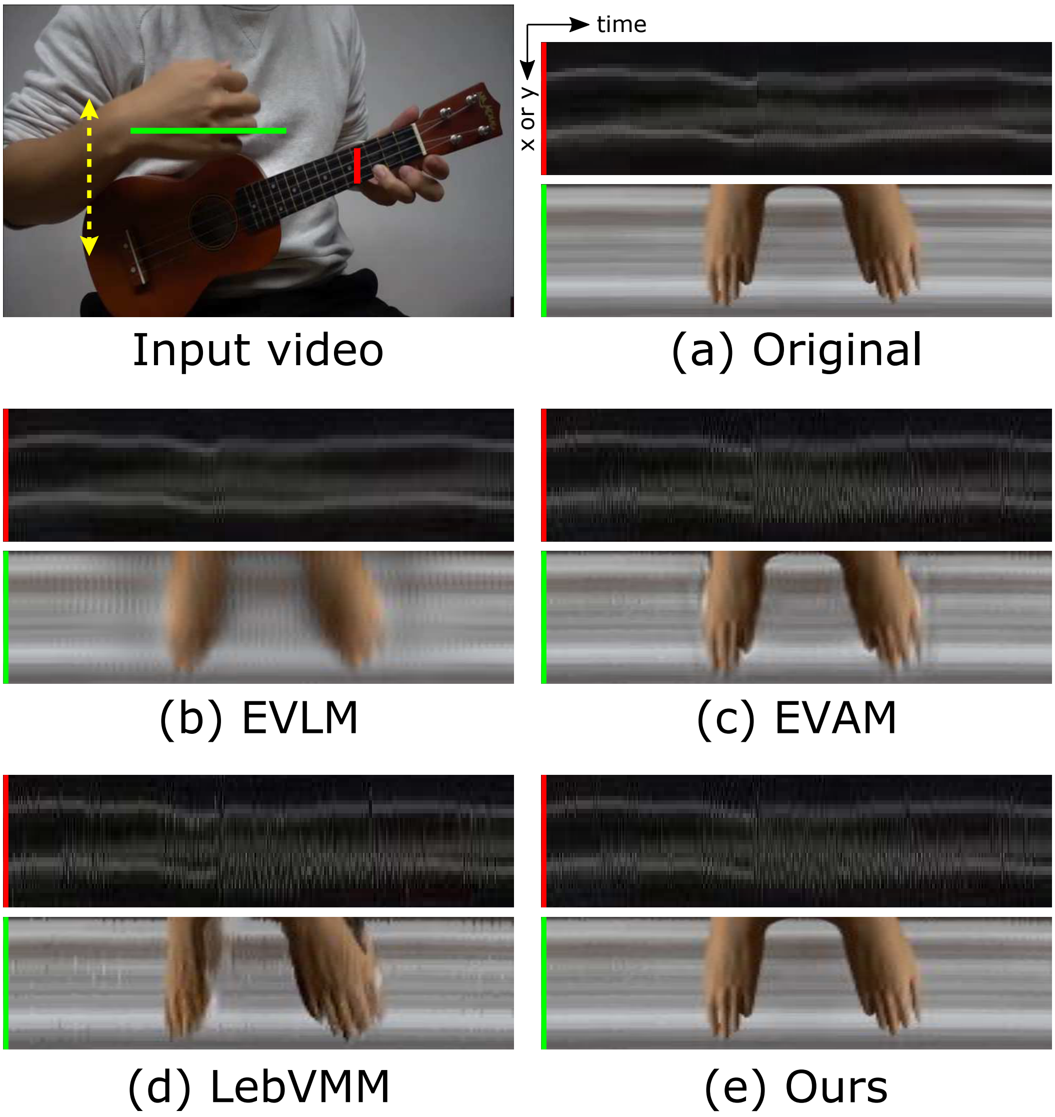


Figure 4.8: Music playing video: ukulele being strummed with repetitive and quick hand motions. Spatiotemporal slices are shown along a single red and green lines (left). Our method automatically ignores all the strumming hand motions and can magnify the subtle vibration of ukulele strings without hand manipulation or additional information.

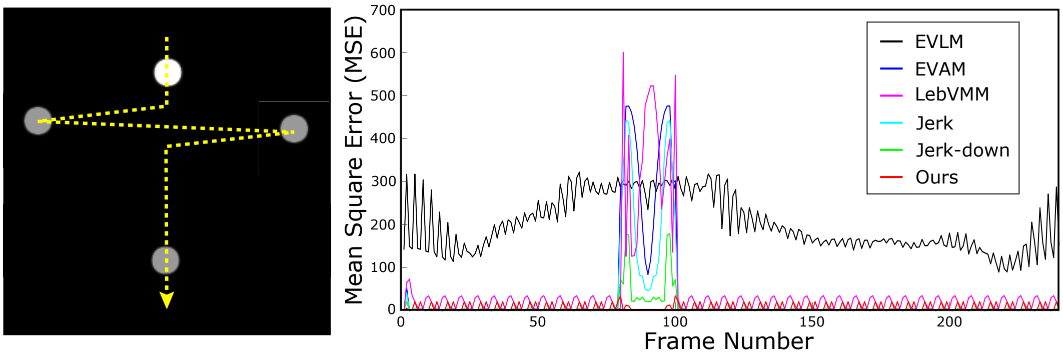


Figure 4.9: Left: synthetic ball video. Four frames are overlaid to indicate ball trajectory depicted with the yellow arrow. The ball exhibits quick large horizontal motion between 80–100 frames. Right: MSE with the ground truth for each frame of this video. Smaller MSE is better. Our method outperforms all other methods.

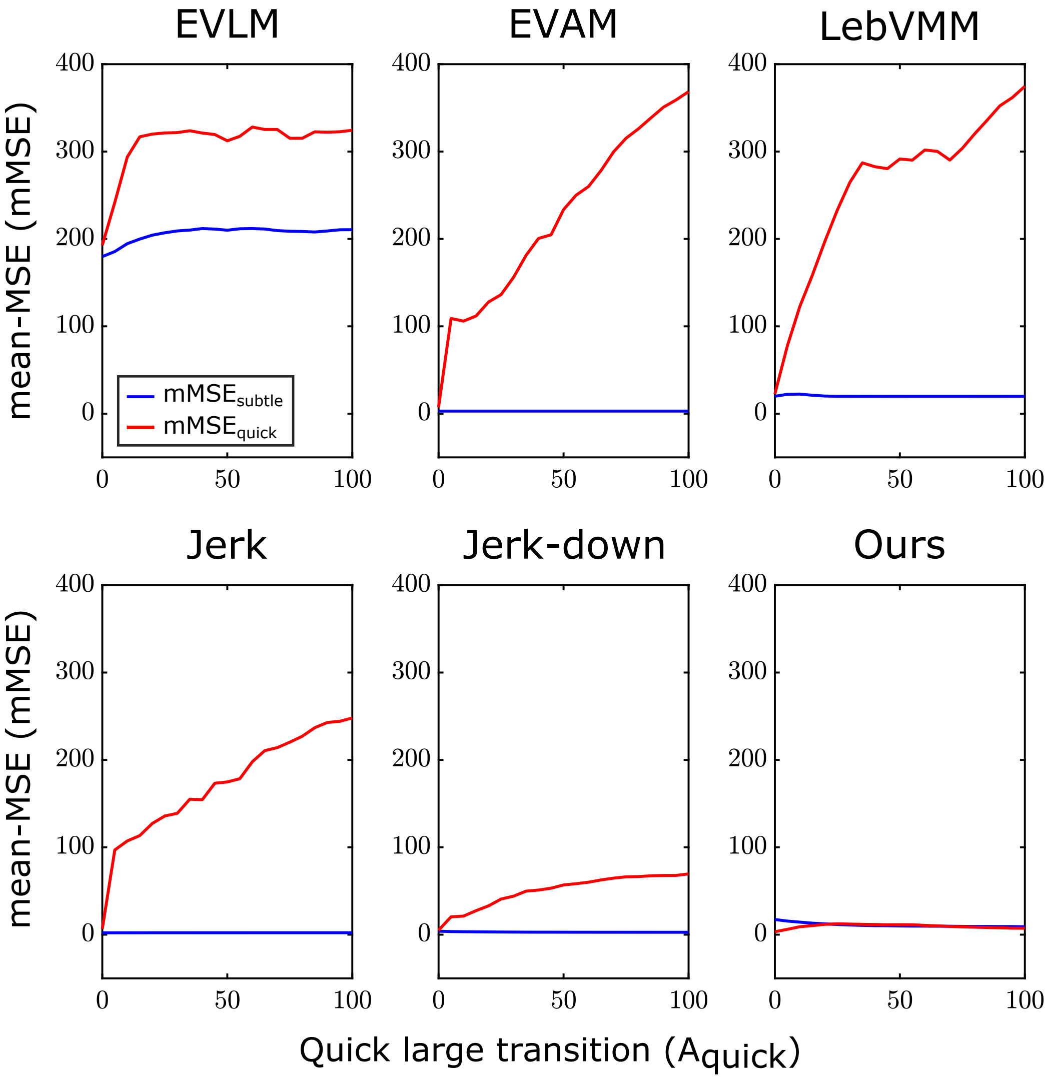


Figure 4.10: Mean-MSE during the appearance of subtle motions in the presence of slow large motions (mMSE$_{\text{subtle}}$) and mean-MSE during the appearance of quick large motions (mMSE$_{\text{quick}}$) with ground truth over different quick large transitions A_{quick}. Our method handles quick large displacement with lower artifacts better than all other methods.

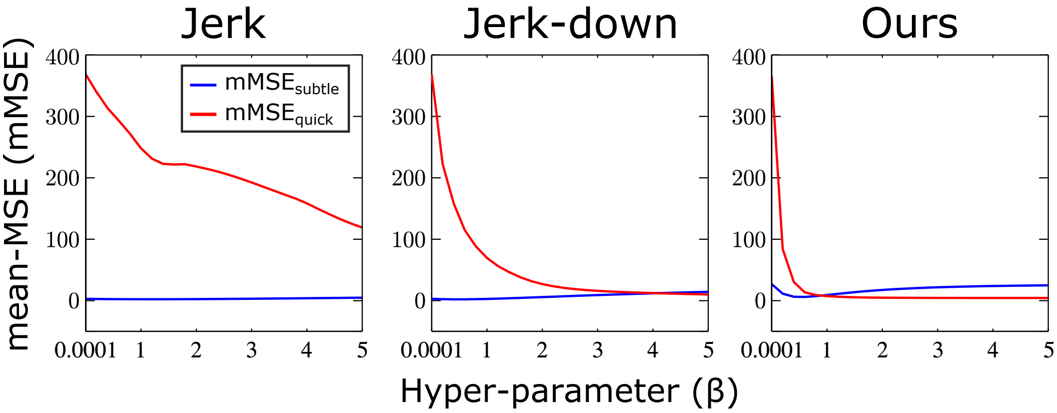


Figure 4.11: mMSE$_{\text{subtle}}$ and mMSE$_{\text{quick}}$ with ground truth over different β. Our proposed method best handles quick large motions while magnifying subtle motions; mMSE$_{\text{quick}}$ is almost 0, and mMSE$_{\text{subtle}}$ is kept low.

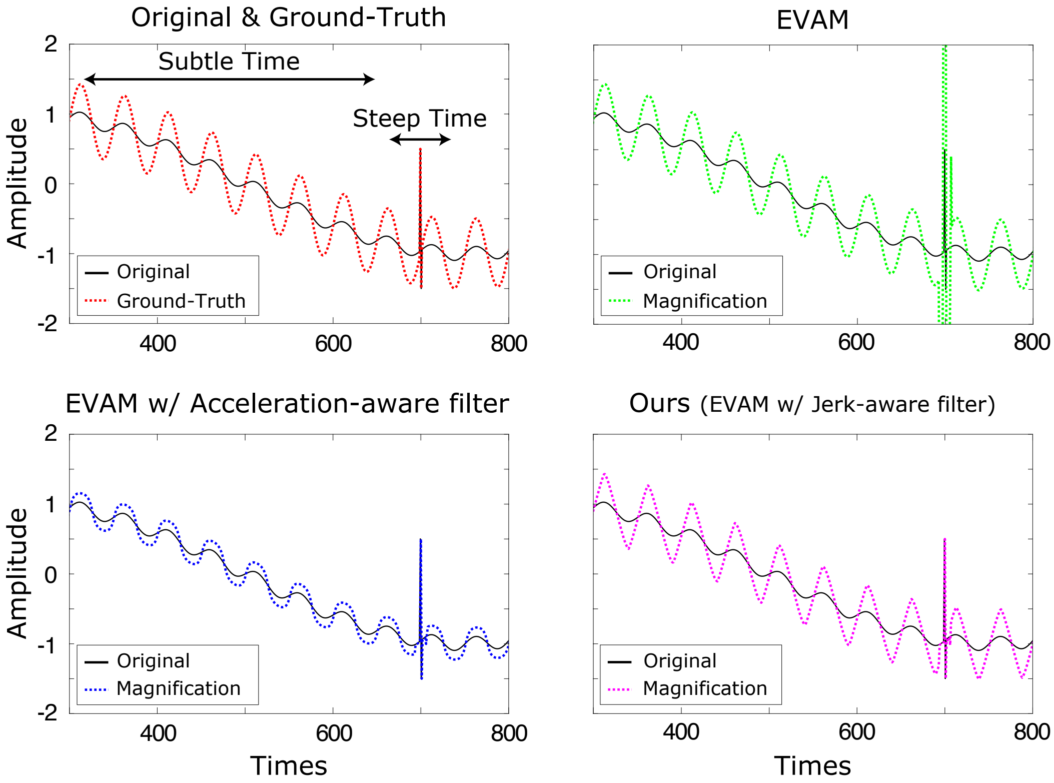


Figure 4.12: Comparisons between our JAF and the acceleration-aware filter in 1D signals. The EVAM method [9] can amplify subtle signals but also accidentally amplifies steep signals (top-right). Applying the acceleration-aware filter makes the EVAM method robust to steep signals but unfortunately suppress subtle signals (bottom-left). On the other hand, our proposed method can suppress only the steep signals and amplify the subtle signals (bottom-right).

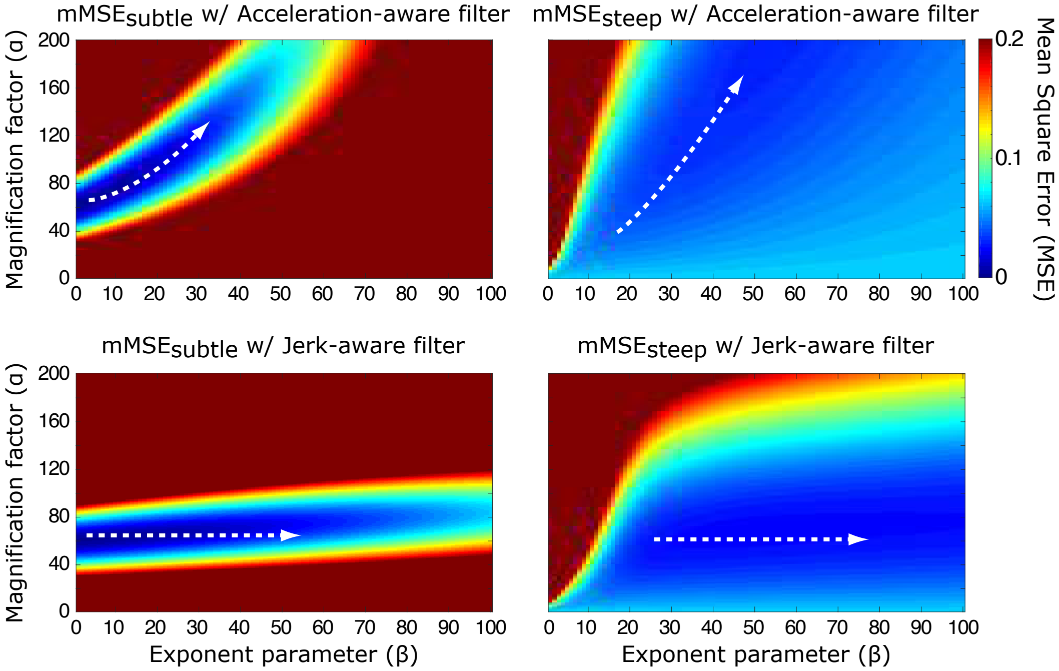


Figure 4.13: Mean-MSEs during the subtle and the steep times at different α and β relative to the ground truth. In the EVAM method [9] with the acceleration-aware filter, α and β strongly correlate in parameter space where mMSEs are low (top row, white arrows). On the other hand, in our proposed method, α and β are independent in parameter space (bottom row, white arrows), so it searches for α during the subtle time and β during the steep time independently to obtain low mMSE.

Chapter 5

Ignoring Photographic Subtle Noise in a video

5.1 Introduction

The conventional EVM methods [4, 5, 9], including our JAEVAM method [48] as explained in Section 4, focus on revealing subtle color changes or motions in the presence of slow and/or quick large motions. However, such subtle changes contain meaningful ones caused by physical/natural phenomena and non-meaningful ones caused by photographic subtle noise. Thus, the conventional EVM methods and our JAEVAM method often produce noisy and misleading results because they mistakenly amplify photographic subtle noise.

For revealing only the meaningful subtle changes in the presence of photographic subtle noise, several methods have been proposed [5, 52, 53]. By focusing on that meaningful motions appear around edges [22], Wadhwa et al. [5] spatially applied an edge-weighted Gaussian filter (EWG) to phase signals at each complex steerable pyramid level, and Verma et al. [52] used a local Laplacian filter (LLP) [54] to improve pyramid decomposition in the LubEVMM method (Subsection 2.3.1) in terms of edges and details. These methods help to remove non-meaningful subtle motions in flat textured regions but have limitations in that they cannot be applied to color magnification or to the removal of non-meaningful subtle motions around edges. Alternatively, Wu et al. [53] adopted PCA to EVM as a pre-processing approach. This method can magnify only meaningful subtle changes in a video, but for enabling PCA to work well, it has a limitation that meaningful subtle changes need to be larger than non-meaningful ones as the principal component in the input video scene.

In this chapter, we propose an EVM method for revealing only meaningful subtle color changes or motions under the presence of photographic subtle noise, without additional interventions, resources, or input video scene limitations. On the basis of our observation that temporal distribution of meaningful subtle changes more clearly indicates anisotropic diffusion than that of non-meaningful ones caused by photographic subtle noise, we considered that the anisotropic diffusion in the temporal distribution enables us to detect only meaningful subtle changes. Therefore, we focused on fractional anisotropy (FA), which is used in neuroscience to evaluate anisotropic diffusion of water molecules in the body for revealing the shape of tiny nerve cells [14, 15]. In developing our method, we used FA to design a fractional anisotropic filter (FAF) that passes only meaningful subtle changes and ignores non-meaningful ones. Additionally, similar to [5, 52], we propose a hierarchical edge-aware regularization (HEAR) for refining uncertain subtle motions at flat (texture-less) regions in a video. Our method, in which FAF and HEAR are applied to the JAEVAM method [48], produces impressive color or motion magnification results in various input video scenes.

5.2 Proposed Method

5.2.1 Problem Definition

Following the preliminary formulation in Chapter 3 and the JAEVAM method [48] described in Chapter 4, given a color/phase signal $S_{\omega_n,\theta}(x,y,t)$, the conventional EVM methods [4, 5, 9, 48] attempt to detect a subtle color/phase signal $C_{\omega_n,\theta,f_t}(x,y,t)$. However, such a subtle signal is often contaminated by photographic subtle noise as

$$C_{\omega_n,\theta,f_t}(x,y,t) = \hat{C}_{\omega_n,\theta,f_t}(x,y,t) + \tilde{C}_{\omega_n,\theta,f_t}(x,y,t), \tag{5.1}$$

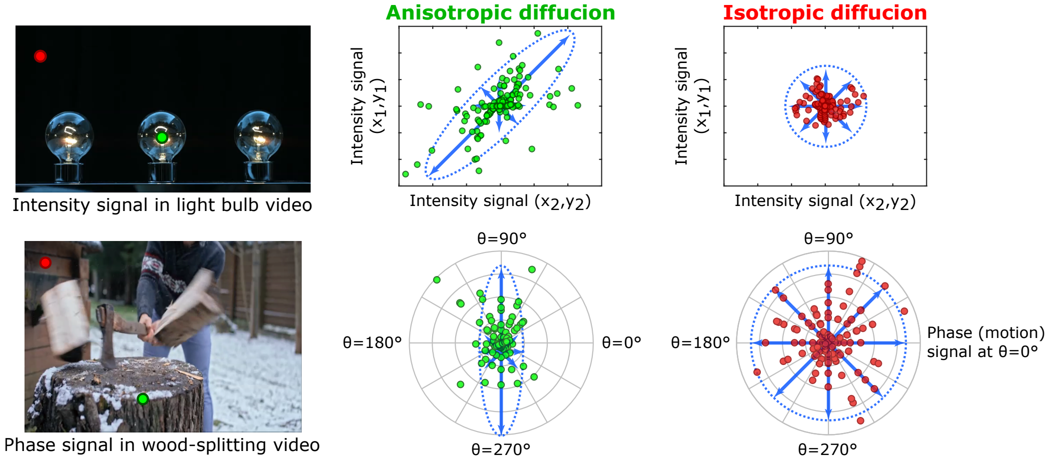
where $\hat{C}_{\omega_n,\theta,f_t}(x,y,t)$ is a meaningful subtle signal and $\tilde{C}_{\omega_n,\theta,f_t}(x,y,t)$ is a non-meaningful one caused by photographic subtle noise. Therefore, the conventional EVM methods often produce noisy and misleading magnification outputs due to $\tilde{C}_{\omega_n,\theta,f_t}(x,y,t)$.

5.2.2 Fractional Anisotropy (FA)

Our key idea is based on our observation that temporal distribution of meaningful subtle changes more clearly indicates anisotropic diffusion than that of non-meaningful ones because they are subject to the regularity of nature (Fig. 5.1). We considered that the anisotropic diffusion in the temporal distribution enables us to detect only meaningful subtle changes and focused on an index called fractional anisotropy (FA).

FA has been used in neuroscience to evaluate anisotropic diffusion of water molecules in the body [14, 15], and its definition is based on the Gaussian diffusion equation as

$$f(\mathbf{g}) = \frac{1}{(2\pi)^{d/2}|\mathbf{D}|^{1/2}} \exp\left(-\frac{1}{2}\mathbf{g}^\top\mathbf{D}^{-1}\mathbf{g}\right) = N(\mathbf{0},\mathbf{D}), \tag{5.2}$$



Intensity signal in light bulb video

Phase signal in wood-splitting video

Figure 5.1: The temporal distributions of subtle luminance color (top) or phase signals that represent motions (bottom). When the meaningful subtle luminance color signals appear, they are correlated between neighboring pixels (top green), but do not when photographic subtle noise only appears (top red). The meaningful subtle phase signals occur in a vertical direction (bottom green) but in no direction if they are not meaningful (bottom red). We noticed that temporal distribution of meaningful subtle changes more clearly indicates anisotropic diffusion than that of non-meaningful ones caused by photographic subtle noise (blue arrow representing the trend).

where $f(\mathbf{g})$ is a probability distribution of water molecules along directions $\mathbf{g} \in \mathbb{R}^d$, and $\mathbf{D}$ is a positive semi-definite matrix that represents diffusion strength of the distribution $f(\mathbf{g})$ along or between directions $\mathbf{g}$. To the best of our knowledge, FA is defined in 3D case ($d = 3$) on the basis of $f(\mathbf{g})$, but we generalize it for multi-dimensional case as

$$\mathrm{FA} := \sqrt{\frac{d}{d-1}} \cdot \frac{\sqrt{\sum_{i=1}^{d} (\lambda_i - \bar{\lambda})^2}}{\sqrt{\sum_{i=1}^{d} \lambda_i^2}}, \tag{5.3}$$

where $(\lambda_1, ..., \lambda_d)$ are eigenvalues of $\mathbf{D}$ and $\bar{\lambda} = \frac{1}{d} \sum_{i=1}^{d} \lambda_i$. The eigenvalues of $\mathbf{D}$

indicate diffusion strength to the direction of eigenvectors in the original directions g. The coefficient $\sqrt{\frac{d}{d-1}}$ normalizes the FA value between 0 and 1. Moreover, we found intuitive interpretation of FA as follows.

Lemma 1.

$$\sum_{i=1}^{d} \left(\lambda_i - \bar{\lambda}\right)^2 = \sum_{i=1}^{d} \lambda_i^2 - d \cdot \bar{\lambda}^2, \tag{5.4}$$

where $\bar{\lambda} = \frac{1}{d} \sum_{i=1}^{d} \lambda_i$.

Proof.

$$\begin{aligned}\sum_{i=1}^{d} \left(\lambda_i - \bar{\lambda}\right)^2 &= \sum_{i=1}^{d} \lambda_i^2 - 2 \cdot \left(\sum_{i=1}^{d} \lambda_i\right) \cdot \bar{\lambda} + d \cdot \bar{\lambda}^2 \\ &= \sum_{i=1}^{d} \lambda_i^2 - 2 \cdot d \cdot \bar{\lambda}^2 + d \cdot \bar{\lambda}^2 \\ &= \sum_{i=1}^{d} \lambda_i^2 - d \cdot \bar{\lambda}^2\end{aligned}$$

□

Lemma 2. Let θ be the angle between two vectors, $(\lambda_1, \ldots, \lambda_d), (1, 1, \ldots, 1) \in \mathbb{R}^d$. Then,

$$\sin\theta = \sqrt{\frac{\sum_{i=1}^{d} \left(\lambda_i - \bar{\lambda}\right)^2}{\sum_{i=1}^{d} \lambda_i^2}}. \tag{5.5}$$

Proof. From the relationship between inner product and norms of two vectors, we get

$$\cos\theta = \frac{\sum_{i=1}^{d} \lambda_i}{\sqrt{\sum_{i=1}^{d} \lambda_i^2} \cdot \sqrt{\sum_{i=1}^{d} 1}} = \frac{\sum_{i=1}^{d} \lambda_i}{\sqrt{\sum_{i=1}^{d} \lambda_i^2} \cdot \sqrt{d}}.$$

Because $\mathbf{D}$ is positive semi-definite and so the vector $(\lambda_1, \ldots, \lambda_d)$ is in the first quadrant, $\sin\theta > 0$ holds. As a result, we can rewrite $\sin\theta$ as follows:

$$\begin{aligned}
\sin\theta &= \sqrt{1 - \cos^2\theta} \\
&= \sqrt{1 - \frac{\left(\sum_{i=1}^{d} \lambda_i\right)^2}{\left(\sum_{i=1}^{d} \lambda_i^2\right) \cdot d}} = \sqrt{1 - \frac{\bar{\lambda}^2 \cdot d^2}{\left(\sum_{i=1}^{d} \lambda_i^2\right) \cdot d}} = \sqrt{\frac{\sum_{i=1}^{d} \lambda_i^2 - \bar{\lambda}^2 \cdot d}{\sum_{i=1}^{d} \lambda_i^2}} \\
&= \sqrt{\frac{\sum_{i=1}^{d} \left(\lambda_i - \bar{\lambda}\right)^2}{\sum_{i=1}^{d} \lambda_i^2}} \quad (\because (5.4)).
\end{aligned}$$

□

Proposition 1.

$$FA = \sqrt{\frac{d}{d-1}} \cdot \sin\theta,$$

where θ is defined in Lemma 2.

Proof. Trivial from Eqs. (5.3) and (5.5). □

This proposition implies that FA purely evaluates the degree of match between the eigenvalues without depending on the magnitude of them. Since the positive semi-definite matrix of $\mathbf{D}$ makes all the eigenvalues positive, if only one eigenvalue is high, which means anisotropic diffusion, θ is maximum and FA value is 1, but if all the eigenvalues are equal, which means isotropic diffusion, θ is 0 and FA value is 0 (Fig. 5.2).

In neuroscience fields, it is known that nerve axons have high FA values due to anisotropic diffusion of water molecules along their long stick structures, but if their axonal structures injury occurs due to such as a traffic accident or a neural disease, the probability distribution of water molecules in the injured area indicates isotropic diffusion and the FA value becomes lower [14, 15]. Thus, a changing of

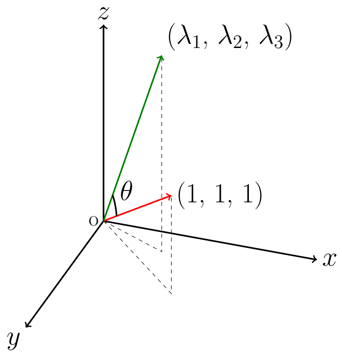


Figure 5.2: Intuitive interpretation of FA in 3D case. FA is proportional to $\sin\theta$, where θ is the angle between two vectors, $(\lambda_1, \lambda_2, \lambda_3)$ which are eigenvalues of $\boldsymbol{D}$, $(1, 1, 1)$. If all the eigenvalues are equal such as $(3, 3, 3)$, which means isotropic diffusion, θ is 0 and the FA value is 0. If only one eigenvalue is high such as $(5, 0, 0)$, which means anisotropic diffusion, θ is maximum and the FA value is 1.

FA value sensitively assesses the loss or recovery process of the shape of nerve cells in humans or animals.

From these findings, we considered that FA values strongly respond to meaningful subtle changes compared with non-meaningful ones, due to the anisotropic diffusion in the temporal distribution of meaningful ones. To visualize our hypothesis, we show FA values estimated by temporal distribution of subtle phase signals in ukulele-playing video (Fig. 5.3). Figure 5.3 indicates that the FA value is higher when the meaningful subtle phase signals appear, such as hand swaying and vibrations of ukulele strings.

5.2.3 Fractional Anisotropic Filter

On the basis of our knowledge of FA, we designed a fractional anisotropic filter (FAF). This filter is designed with FA estimated from diffusion in temporal distribution of subtle color/phase signals so that it will pass only meaningful subtle color changes or motions, which have high FA values. First, we get FA value as follows.

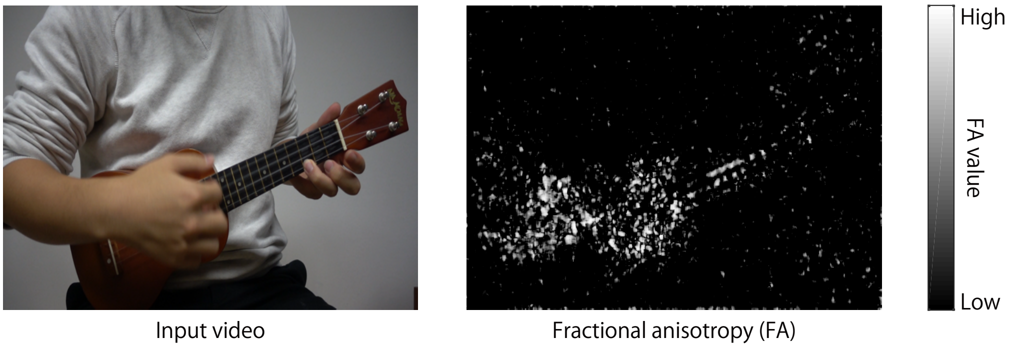


Figure 5.3: Visualizing fractional anisotropy (FA) values. FA values are estimated by temporal distribution of subtle phase changes and are high when the meaningful subtle phase changes appear such as hand swaying and vibrations of ukulele strings.

For Color Magnification

The subtle color signal $C_{\omega_n,\theta,f_t}(x, y, t)$ can be written as $C_{n,f_t}(x, y, t)$ following $\theta = \emptyset$ as explained in Chapter 3. Given a $(h \times w)$-dimensional vector $\mathbf{g}_{n,f_t}(x, y, t) = [C_{n,f_t}(x_1, y_1, t), \ldots, C_{n,f_t}(x_{h\times w}, y_{h\times w}, t)]^\top$ that represents subtle color signals within $(x_i, y_i) \in \mathcal{N}(x, y)$, where $\mathcal{N}(x, y)$ indicates the neighborhood image positions of (x, y), we assume a set of vectors $\{\mathbf{g}_{n,f_t}(x, y, t_k) \mid t_k \in \mathcal{N}(t)\}$, where $\mathcal{N}(t)$ indicates the neighborhood time frames of t, is created by i.i.d sampling from a temporal distribution $f(\mathbf{g}_{n,f_t}(x, y, t))$ defined as

$$f(\mathbf{g}_{n,f_t}(x, y, t)) = N\left(\mathbf{0}, \mathbf{D}_{n,f_t}(x, y, t)\right). \tag{5.6}$$

Using maximum likelihood estimation method, we estimate $\mathbf{D}_{n,f_t}(x, y, t)$ representing diffusion strength of the temporal distribution $f(\mathbf{g}_{n,f_t}(x, y, t))$ along or between the image positions in $\mathcal{N}(x, y)$ as

$$\mathbf{D}_{n,f_t}(x, y, t) = \frac{1}{|\mathcal{N}(t)|} \sum_{t_k \in \mathcal{N}(t)} (\mathbf{g}_{n,f_t}(x, y, t_k) - \bar{\mathbf{g}})(\mathbf{g}_{n,f_t}(x, y, t_k) - \bar{\mathbf{g}})^\top, \tag{5.7}$$

where $\bar{\mathbf{g}} = \frac{1}{|\mathcal{N}(t)|} \sum_{t_k \in \mathcal{N}(t)} \mathbf{g}_{n,f_t}(x, y, t_k)$. Then, we get $\mathrm{FA}_{n,f_t}(x, y, t)$ by using Eq. (5.3) in the case of Eq. (5.7).

After calculating $\mathrm{FA}_{n,f_t}(x, y, t)$, we design the fractional anisotropic filter $\mathrm{FAF}_{n,f_t}(x, y, t; \sigma_s, \gamma)$ via min-max normalization function $\Gamma(\cdot)$ with a 2D Gaussian smoothing with a variance σ_s^2 and a weight γ for adjusting the filter response as

$$\mathrm{FAF}_{n,f_t}(x, y, t; \sigma_s, \gamma) = \Gamma\left(G_{\sigma_s}(x, y) * \mathrm{FA}_{n,f_t}(x, y, t)\right)^{\gamma}, \tag{5.8}$$

where $G_{\sigma_s}(x, y)$ is a 2D Gaussian filter with σ_s^2 to smooth the filter responses. This filter has a high value only when anisotropic diffusion in temporal distribution of subtle color signals appears, which means it can pass only meaningful subtle color changes and ignores non-meaningful ones.

For Motion Magnification

Given a M-dimensional vector $\mathbf{g}_{\omega_n,f_t}(x, y, t) = [C_{\omega_n,\theta_1,f_t}(x, y, t), \ldots, C_{\omega_n,\theta_M,f_t}(x, y, t)]^\top$ that represents subtle phase signals at orientations $\{\theta_1, \ldots, \theta_M\} \in \Theta$, we assume a set of vectors $\{\mathbf{g}_{\omega_n,f_t}(x_i, y_j, t_k) \mid (x_i, y_j) \in \mathcal{N}(x, y), t_k \in \mathcal{N}(t)\}$ is created by i.i.d sampling from a temporal distribution $f(\mathbf{g}_{\omega_n,f_t}(x, y, t))$ defined as

$$f(\mathbf{g}_{\omega_n,f_t}(x, y, t)) = N\left(\mathbf{0}, \mathbf{D}_{\omega_n,f_t}(x, y, t)\right). \tag{5.9}$$

Using maximum likelihood estimation method, we estimate $\mathbf{D}_{\omega_n,f_t}(x, y, t)$ representing diffusion strength of the temporal distribution $f(\mathbf{g}_{\omega_n,f_t}(x, y, t))$ along or between the orientations $\theta_1, \ldots, \theta_M$ as

$$\begin{aligned}&\mathbf{D}_{\omega_n,f_t}(x, y, t)\\ &= \frac{1}{|\mathcal{N}(x, y)||\mathcal{N}(t)|} \sum_{\substack{(x_i,y_j)\in\mathcal{N}(x,y)\\ t_k\in\mathcal{N}(t)}} (\mathbf{g}_{\omega_n,f_t}(x_i, y_j, t_k) - \bar{\mathbf{g}})(\mathbf{g}_{\omega_n,f_t}(x_i, y_j, t_k) - \bar{\mathbf{g}})^\top,\end{aligned} \tag{5.10}$$

where $\bar{\mathbf{g}} = \frac{1}{|\mathcal{N}(x,y)||\mathcal{N}(t)|} \sum_{(x_i,y_j)\in\mathcal{N}(x,y), t_k\in\mathcal{N}(t)} \mathbf{g}_{\omega_n,f_t}(x_i, y_j, t_k)$. Then, we get $\mathrm{FA}_{\omega_n,f_t}(x, y, t)$ by using Eq. (5.3) in the case of Eq. (5.10).

After calculating $\text{FA}_{\omega_n,f_t}(x,y,t)$, we design the fractional anisotropic filter $\text{FAF}_{\omega_n,f_t}(x,y,t;\sigma_s,\gamma)$ via min-max normalization function $\Gamma(\cdot)$ with a 2D Gaussian smoothing with a variance σ_s^2 and a weight γ for adjusting the filter response as

$$\text{FAF}_{\omega_n,f_t}(x,y,t;\sigma_s,\gamma) = \Gamma\left(G_{\sigma_s}(x,y) * \text{FA}_{\omega_n,f_t}(x,y,t)\right)^{\gamma}, \tag{5.11}$$

where $G_{\sigma_s}(x,y)$ is the same of Eq. (5.8). This filter has a high value only when anisotropic diffusion in temporal distribution of subtle phase signals appears, which means it can pass only meaningful subtle motions and ignores non-meaningful ones.

5.2.4 Hierarchical Edge-Aware Regularization (HEAR)

For refining subtle phase signals (not color signals), we focus on a subband amplitude signal $A_{\omega_n,\theta}(x,y,t)$ in the same way as EWG of the previous method [5]. However, while the previous method directly uses a subband amplitude signal at each pyramid level n, we develop hierarchical amplitude correction via z-transform as

$$\hat{A}_{\omega_n,\theta}(x,y,t) = \max_{i \in \mathcal{N}(n)} \left(\mathcal{Z}\left(A_{\omega_n,\theta}(x,y,t)\right), \mathcal{R}\left(\mathcal{Z}\left(A_{\omega_i,\theta}(x,y,t)\right), n\right)\right), \tag{5.12}$$

where $\mathcal{N}(n)$ indicates the neighborhood pyramid levels of n, $\mathcal{R}(\cdot,n)$ is the same of Eq. (4.8), and $\mathcal{Z}(\cdot)$ converts the input into z-score that are comparable between each n.

Furthermore, as the previous method [5] adopts amplitude-based smoothing but is weak in regularizing flat textured areas due to its use of smoothing alone, we propose a hierarchical edge-aware regularization $\text{HEAR}_{\omega_n,\theta}(x,y,t)$ via min-max normalization function $\Gamma(\cdot)$ with a 2D Gaussian smoothing with a variance σ_s^2 as

$$\text{HEAR}_{\omega_n,\theta}(x,y,t;\sigma_s) = \Gamma\left(G_{\sigma_s}(x,y) * \hat{A}_{\omega_n,\theta}(x,y,t)\right). \tag{5.13}$$

This form is similar to the design of FAF in Eqs. (5.8) and (5.11).

5.2.5 Applying FAF and HEAR to the JAEVAM method

We apply FAF of Eq. (5.8) and JAF of Eq. (4.7) to Eq. (4.3) for color magnification. On the other hand, we apply FAF of Eq. (5.11), HEAR of Eq. (5.13), and the propagated JAF of Eq. (4.8) to Eq. (4.3) for motion magnification. Through these applications, we obtain a result where only meaningful subtle color changes or motions are magnified under the presence of slow and quick large motions.

Figure 5.4 shows the effect of our method on detecting phase signals in an ukulele-playing video. (a,b) The EVAM method [9] and the JAEVAM method [48] cannot ignore the non-meaningful subtle phase signals. By applying FAF to the JAEVAM method [48], we can suppress the non-meaningful subtle phase signals but slightly misdetects them in flat textured areas (c, purple quadrangles). By further applying HEAR to (c), we eventually can detect only meaningful subtle phase signals of the ukulele strings.

5.3 Results

5.3.1 Experimental Setup

To evaluate the effectiveness of our method, we conducted experiments on real videos and synthetic ones with ground-truth magnification. We assessed the effectiveness qualitatively for real videos and quantitatively against ground-truth for synthetic ones. We set the parameters for each experiment as listed in Table 5.1, and σ_s in Eqs. (5.8), (5.11), and (5.13) is equal to the spatial filter widths used to construct the Gaussian or the complex steerable pyramid at each pyramid level n.

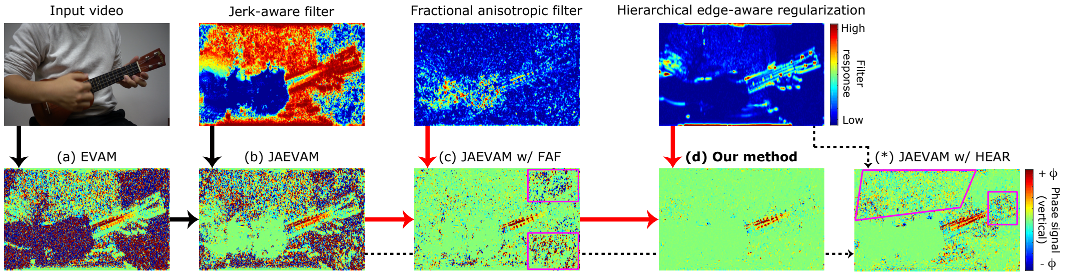


Figure 5.4: Our motion magnification method. (a) The EVAM method [9] misdetects quick hand strumming. (b) The JAEVAM method [48] ignores the quick motion but misdetects non-meaningful subtle phase signals caused by photographic noise. (c) By applying FAF to the JAEVAM method [48], we can suppress the non-meaningful subtle phase signals but slightly misdetects them in flat textured areas (purple quadrangles). (d) Our method further applies HEAR to refine them and we eventually can detect only meaningful subtle phase signals of the ukulele strings. (*) Using only HEAR is insufficient for complex areas (purple quadrangles). These results indicate both of FAF and HEAR are needed.

Color Magnification

We constructed a Gaussian pyramid to decompose each image frame into multi-scales and magnified the G color signals on the fifth pyramid level.

Motion Magnification

We performed each method in Y color channel. To obtain subband amplitude and subband phase signals from input video, we constructed a complex steerable pyramid with half-octave bandwidth filters and 8 orientations. We set parameter $V = 5$ in the JAEVAM method [48] and $|\mathcal{N}(n)| = 5$ in Eq. (5.12). For designing FAF (Eq. (5.8) or Eq. (5.11)), we set $|\mathcal{N}(x, y)| = (5 \times 5)$, and $|\mathcal{N}(t)|$ as the same time length used to detect subtle signals at f_t in TAF [9].

Table 5.1: Parameters for all videos: amplification factor α in our method (this parameter in other methods was adjusted to magnify meaningful subtle changes as much as ours), target temporal frequency f_t, sampling frequency f_s, large motions suppression parameter β in the JAEVAM method [48], and hyper parameter γ in Eq. (5.8) or Eq. (5.11).

Video	α	f_t	f_s	β	γ	source
Slam dunk	200	2	120	1	2	[55]
Ukulele	260	40	240	1	5	[48]
Face	180	0.5	60	0.001	3	[4]
Wood	230	2	120	3	2	[55]
Gun	100	20	480	0.5	1	[48]
Tennis	180	10	600	1	1	[55]

5.3.2 Real Videos

We compared our proposed method with two state-of-the-art methods, The EVAM method [9] and the JAEVAM method [48], both of which can perform color or motion magnification without user annotations or additional information in the same way as our method.

Comparison with Color Magnification

Figure 5.5 illustrates subtle face color changes due to blood flow through the face of a stationary man. Processing this video with the EVAM method [9] or the JAEVAM method [48] succeeds in magnifying meaningful subtle face color changes on the face, but it also misdetects and magnifies non-meaningful background color fluctuations caused by photographic noise. In contrast, our proposed method magnifies only meaningful subtle face color changes.

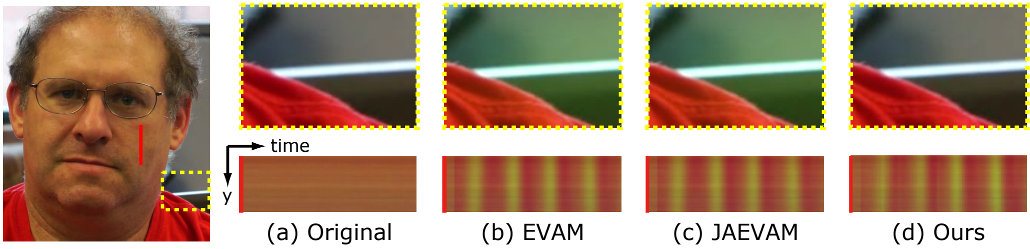


Figure 5.5: Color magnification at blood flow through the face of a stationary man. Our proposed method magnifies only meaningful subtle face color changes (bottom), while the EVAM method [9] and the JAEVAM method [48] misdetect and magnify non-meaningful background color fluctuations caused by photographic noise (top).

Comparison with Motion Magnification

Figure 5.6 shows the motion magnification results from a basketball video, to magnify and reveal the subtle deformations of the backboard when trying to absorb the impact of a slam dunk for preventing breakage. The EVAM method [9] does not work well due to the misdetection of the quick ball motion. The JAEVAM method [48] magnifies meaningful subtle deformation of the backboard but also misdetects non-meaningful subtle shape collapses of background window caused by photographic noise. In contrast, our proposed method magnifies only meaningful subtle deformations of the backboard without the effects of noise.

Figure 5.7 shows a video sequence on the ability of a wood-splitting stand to absorb the shock from a hand axe for preventing injury. The EVAM method [9] produces messy result due to the quick downswing of the hand axe. The JAEVAM method [48] can magnify subtle deformations of the wood-splitting stand but produces pixel intensity disturbances due to non-meaningful background fluctuations caused by photographic noise. Our method magnifies only meaningful subtle deformations of the wood-splitting stand under the presence of photographic noise.

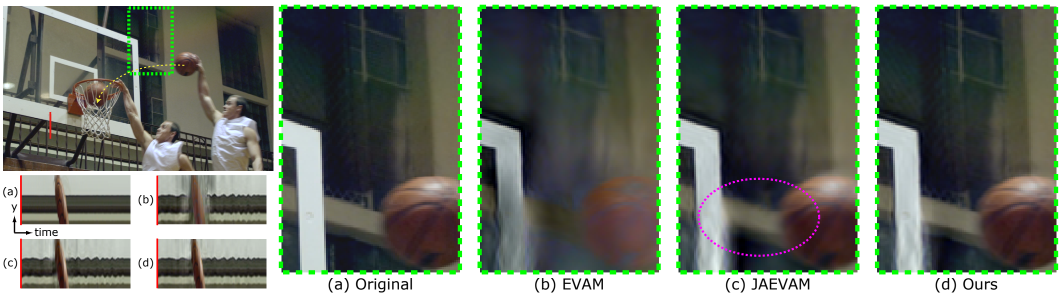


Figure 5.6: Top left: slam dunk video visualizing backboard deformations with ball trajectory by yellow arrow. Bottom left (a)-(d) show spatio-temporal slices along the single red line at top left. Right (a)-(d) show backgrounds in the green square at top left. (b) The EVAM method [9] produces messy artifacts due to quick ball motion. (c) The JAEVAM method [48] magnifies meaningful subtle backboard deformations but misdetects non-meaningful subtle distortions of background window caused by photographic noise (purple circle). (d) On the contrary, our proposed method magnifies only meaningful subtle backboard deformations. See supplementary material for video results.

Figure 5.8 shows a gun-shooting video. In this video, we also tested the LebVMM method [27] with a $5\times$ dynamic mode. The JAEVAM method [48] misdetects distortions of background caused by photographic noise. The LebVMM method [27] also misdetects them slightly and induces disappearance of the tip of the gun due to quick gun recoil motions. Our method magnifies only meaningful subtle deformations of muscles and skin due to the gun-shooting impact spreading throughout the body.

Figure 5.9 shows a ball-hitting video with magnification of impact spreading throughout a tennis racket. The EVAM method [9] produces racket shape collapse due to the quick swing motion. The JAEVAM method [48] magnifies subtle racket deformations when the ball is hit but induces pixel intensity disturbances due to non-

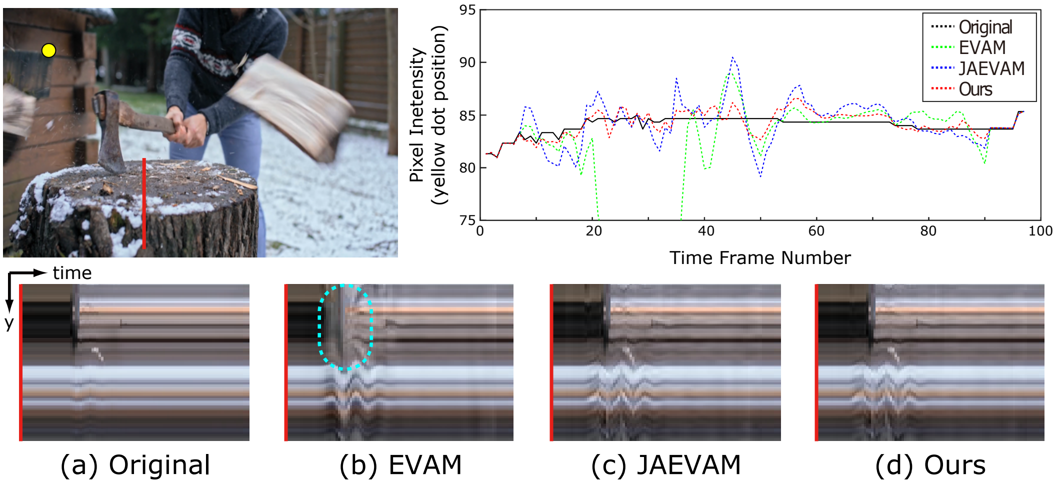


Figure 5.7: Wood-splitting video: visualizing deformations of a wood-splitting stand. The graph shows pixel intensity changes at yellow dot in top left. Our proposed method magnifies only meaningful subtle deformations of the wood-splitting stand, while the EVAM method [9] misdetects the quick downswing of hand axe (cyan circle) and the JAEVAM method [48] produces pixel intensity disturbance due to non-meaningful background fluctuations (graph).

meaningful background fluctuations caused by photographic noise. In contrast, our method magnifies only meaningful deformations related to sport activities under the presence of photographic noise.

5.3.3 Controlled Experiments

In this section, we quantitatively assess the effectiveness of our method using peak signal-to-noise ratio (PSNR) between magnified synthetic video by each magnification method and the ground-truth. Figure 5.10 (top left) shows a 4-second synthetic ball video with background texture from the Describable Textures Dataset [56]. The ball has vertical **meaningful** subtle motions defined as $d = 0.5 \cdot \sin(2\pi \frac{f_t}{f_s} j)$, where j is the frame number. When j reaches 80 frames, the ball moves quickly

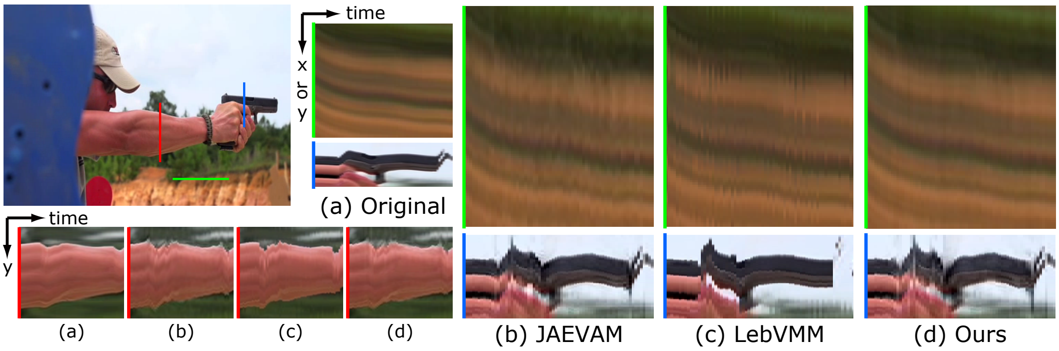
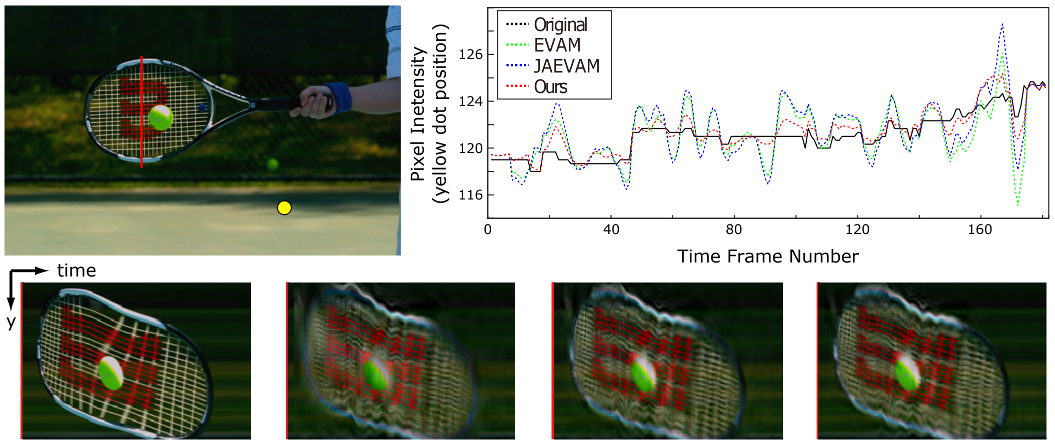


Figure 5.8: Gun-shooting video: visualizing gun-shooting impact spreading throughout body. Our method magnifies only meaningful subtle arm deformations (left bottom) but the JAEVAM method [48] misdetects background distortions caused by photographic noise (right top) and the LebVMM method [27] induces disappearance of the tip of the gun due to quick gun recoil motions (right bottom).

and horizontally as $d_q = 100 \cdot \sin(2\pi \frac{2}{f_s} j)$, but after 20 frames the ball movement returns to what it was before. Moreover, Gaussian noise with an average of 0 and standard deviation σ_n of 0–0.1 was added to only the background in a videos as the photographic noise that causes **non-meaningful** subtle motions. To obtain the ground-truth of meaningful subtle motion magnification, we created magnification videos while changing d to $5 \cdot d$.

Note that to investigate the effectiveness of our proposed method precisely, we prepared five additional methods: a JAEVAM method with EWG proposed by [5], a JAEVAM method with PCA, a JAEVAM method with FAF, a JAEVAM method with No-hierarchical edge-aware regularization as $\mathrm{EAR}_{\omega_n,\theta}(x, y, t; \sigma_s) = \Gamma\left(G_{\sigma_s}(x, y) * A_{\omega_n,\theta}(x, y, t)\right)$, and a JAEVAM method with HEAR.

Figure 5.10 right shows PSNR in each area and each background, at the real noise level $\sigma_n = 0.005$ estimated by [57]. In the ball area, the LubEVMM methods [52, 53], the EVAM method [9] and the LebVMM method [27] suffer from handling quick motion and produce low PSNR, but all JAEVAM based methods, which



(a) Original (b) EVAM (c) JAEVAM (d) Ours

Figure 5.9: Tennis video: visualizing impact spreading throughout a tennis racket. Our method magnifies only meaningful subtle tennis racket deformations, but the EVAM method [9] and the JAEVAM method [48] produce pixel intensity disturbance due to non-meaningful background fluctuations caused by photographic noise (graph).

contain our method, magnify only meaningful subtle motions and have high PSNR except for the JAEVAM method with PCA, which cannot magnify meaningful ones due to large non-meaningful ones regarded as a principal component. On the other hand, in the noise area, the JAEVAM method [48] produces very low PSNR due to non-meaningful subtle motions magnified by the large amplification factor compared with the EVAM method [9]. The JAEVAM method with EWG [5], PCA, our proposed FAF, and HEAR ignore non-meaningful ones and increase PSNR compared with the JAEVAM method [48] but all of these are insufficient. Our proposed method, which considers anisotropic diffusion in temporal distribution by FAF and hierarchical amplitude information by HEAR, ignores non-meaningful ones very well and has high PSNR in the noise area. After all, our proposed method magnifies only meaningful subtle ball motions under the presence of noise and has the

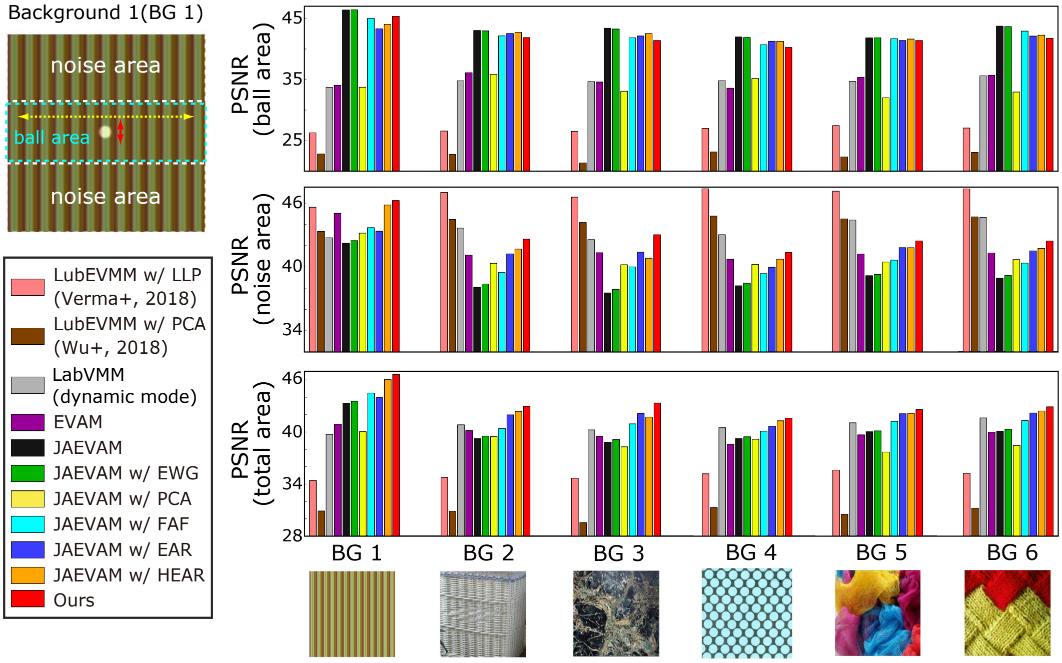


Figure 5.10: Left: synthetic ball video with background. The ball has **meaningful** subtle motions (red arrow) and quick motions (yellow arrow). Noise is added only to background and causes **non-meaningful** subtle motions. Right: PSNR at $\sigma_n = 0.005$. Our proposed method magnifies only meaningful subtle ball motions under the presence of noise and has the highest PSNR in the total area despite the complex background textures.

highest PSNR in the total area despite the complex background textures.

Figure 5.11 shows the effect of noise variance σ_n on the average of PSNR for all the background videos. In the ball area, each magnification method maintains almost the same PSNR. However, the JAEVAM method with PCA cannot do so because the principal component in a video is switched from meaningful subtle motions to non-meaningful ones when $\sigma_n = 0.005$. In the noise area, PSNR in all methods gets lower in proportion to the noise increase. However, if we compare each magnification method for the total area, our proposed method resists

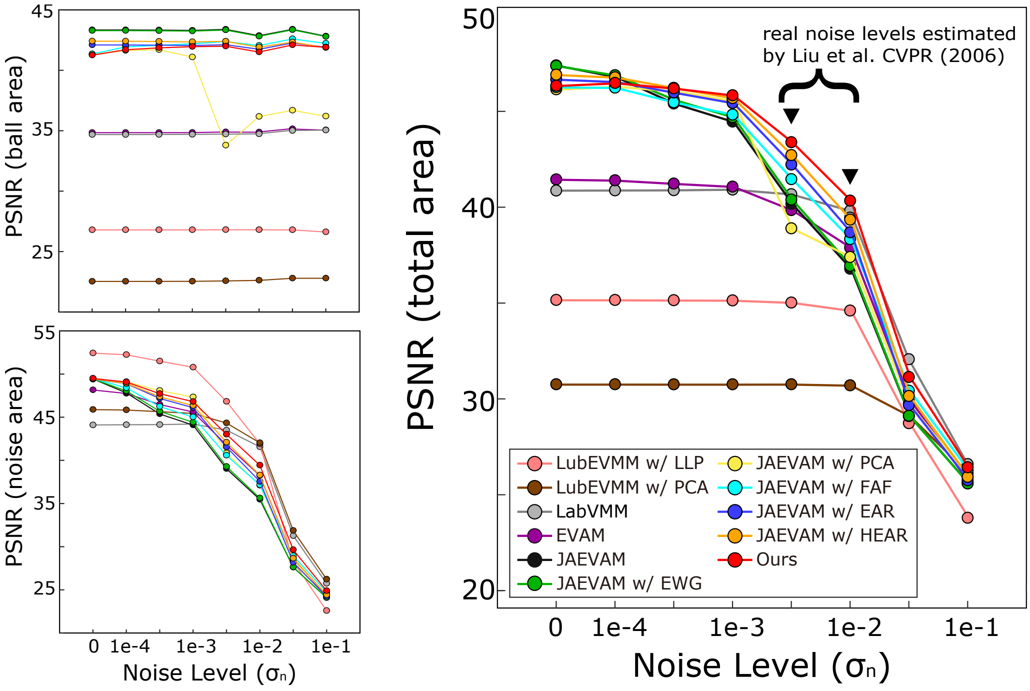


Figure 5.11: The effect of noise variance σ_n on PSNR for all the background videos on average. In the total area, our proposed method resists noise increase and has the highest PSNR in the real noise situations [57].

the effect of noise increase and has the highest PSNR in the real noise situations ($\sigma_n = 0.005, 0.01$). Thus, our method produces the best meaningful and non-misleading magnification results.

5.4 Discussions and Limitations

Our method expands the applicable range of EVM by revealing only meaningful subtle color changes or motions under the presence of photographic subtle noise but has some limitations below.

Our proposed FAF can detect only meaningful subtle changes, but it relies on the assumption that the temporal distribution of non-meaningful ones caused by photographic subtle noise indicates isotropic diffusion. In real videos, such a characteristic like Gaussian distribution often occurs but other ones also need to be considered: gamma, exponential, uniform, impulse, and so on [58]. Thus, we should handle such characteristics to expand the applicable range of video magnification in future work.

If an input video size is large, our method has slow running time due to the eigen-decomposition at each position, time, and pyramid level after. If one wants to precisely reveal meaningful subtle changes and show the results, our method should be used to prevent magnified non-meaningful changes that may be misleading. However, a faster algorithm for our method needs to be developed.

Moreover, empirical estimation of covariance in FA of our method (Eq. (5.7) and Eq. (5.10)) is not robust to outliers under the Gaussian assumption in Eq. (5.3). To increase the robustness, we consider that a minimum covariance determinant approach [59] can be useful. Even so, we should develop a simple and principled approach as a substitute for using FA in future work.

Chapter 6

Accelerating Computational Time

6.1 Introduction

As explained in Section 2.3 and Chapter 3, conventional EVM methods construct over-complete image pyramid representations when analyzing subtle color/phase signals in a video. In EVCM, as subtle color signals are analyzed at only a certain Gaussian pyramid level, its computational time is not big of a deal. However in PbEVMM, subtle phase signals representing local motions are analyzed at all complex steerable pyramid levels and steerable orientations. Therefore, PbEVMM methods require a long computational time in proportion to video resolution and time frame length, compared with EVCM.

For accelerating the computational time of PbEVMM, Wadhwa et al. [6] have proposed a Riesz pyramid as an improvement to the complex steerable pyramid used in previous methods [5, 7, 8, 9, 48, 51]. In the complex steerable pyramid, Hilbert transform is performed along each steerable orientation to detect phase signals at each steerable orientation. On the other hand, in the Riesz pyramid, Riesz transform, which generalizes Hilbert transform in a multi-dimensional manner, is performed to only detect phase signals at the dominant orientation; the Riesz pyramid thus succeeds in removing arbitrariness of orientations in the complex steerable pyramid. Therefore, the Riesz pyramid lowers the over-completeness of the complex steerable pyramid and enables us to analyze subtle phase signals faster and to amplify them quickly. However, since the entire image frames must be processed, the construction using the Riesz pyramid still requires a long computational time in proportion to the video resolution and time frame length.

In this chapter, we propose a faster PbEVMM method using a sophisticated image pyramid called a local Riesz pyramid. The local Riesz pyramid newly adopts local image processing when analyzing subtle phase signals in the Riesz pyramid. We considered that we only have to process the minimum number of sufficient local image areas at a pyramid level n related to the strongly magnified image areas at the above $n+1$ because phase signals have a correlation between adjacent pyramid levels as reported in [16, 17]. From this consideration, our proposed method with the local Riesz pyramid analyzes phase signals at $n+1$ as those with the Riesz pyramid [6] and then detect strongly amplified image areas by using Otsu's thresholding method [60]. After this detection, the strongly amplified local image areas at $n+1$ are propagated to the below n, and then we amplify subtle phase signals in only those areas at n. Our PbEVMM method with the local Riesz pyramid produces impressive motion magnification results equivalent to conventional methods within a short computational time in both real and synthetic videos.

6.2 Conventional Method: Riesz Pyramid for Fast Phase-based Eulerian Video Motion Magnification

Here, we explain a conventional fast PbEVMM method using Riesz pyramid proposed by Wadhwa et al. [6].

Given a normalized image signal $I(x, y, t) \in [0, 1]$ at Y color channel, they first construct a non-oriented subbands pyramid, e.g. a Laplacian pyramid, $\{L_{\omega_n}(x, y, t) \mid n = 0, \ldots, N-1\}$, where $L_{\omega_n}(x, y, t) \in \mathbb{R}^{H_n \times W_n}$ is a subband image signal with an image height H_n and an image width W_n. Next, the Riesz transform, which generalizes a one-dimensional Hilbert transform into a multi-dimensional one [61], is applied to $L_{\omega_n}(x, y, t)$ as follows.

$$\begin{aligned} L_{\omega_n}(x, y, t) &\xrightarrow{\mathscr{F}} \mathfrak{L}_{\omega_n}(\omega_x, \omega_y, t), \\ R^1_{\omega_n}(x, y, t) \xleftarrow{\mathscr{F}^{-1}} \mathfrak{R}^1_{\omega_n}(\omega_x, \omega_y, t) &= \mathfrak{L}_{\omega_n}(\omega_x, \omega_y, t) \cdot -i\frac{\omega_x}{\sqrt{\omega_x^2 + \omega_y^2}}, \\ R^2_{\omega_n}(x, y, t) \xleftarrow{\mathscr{F}^{-1}} \mathfrak{R}^2_{\omega_n}(\omega_x, \omega_y, t) &= \mathfrak{L}_{\omega_n}(\omega_x, \omega_y, t) \cdot -i\frac{\omega_y}{\sqrt{\omega_x^2 + \omega_y^2}}, \end{aligned} \tag{6.1}$$

where $R^1_{\omega_n}(x, y, t)$ and $R^2_{\omega_n}(x, y, t)$ are Riesz-transformed subband image signals along x- and y-axis respectively, $\mathscr{F} : \mathbb{R}^{H_n \times W_n} \to \mathbb{C}^{H_n \times W_n}$ is the 2D Fourier transform, $\mathscr{F}^{-1} : \mathbb{C}^{H_n \times W_n} \to \mathbb{R}^{H_n \times W_n}$ is the inverse 2D Fourier transform, and $\omega_x, \omega_y \in [-\pi, \pi]$ represents the x- and y-axis spatial angular frequency at the pyramid level n. Through this Riesz transform, we can construct a Riesz pyramid consisting of a set of three subband image signals: a subband image signal $L_{\omega_n}(x, y, t)$ and two Riesz-transformed subband image signals $R^1_{\omega_n}(x, y, t)$ and $R^2_{\omega_n}(x, y, t)$ as shown in Fig. 6.1.

Additionally, Wadhawa et al. [6] pointed out a long computational time of the 2D Fourier transform $\mathscr{F}$ and the inverse 2D Fourier transform $\mathscr{F}^{-1}$ in Eq. (6.1)

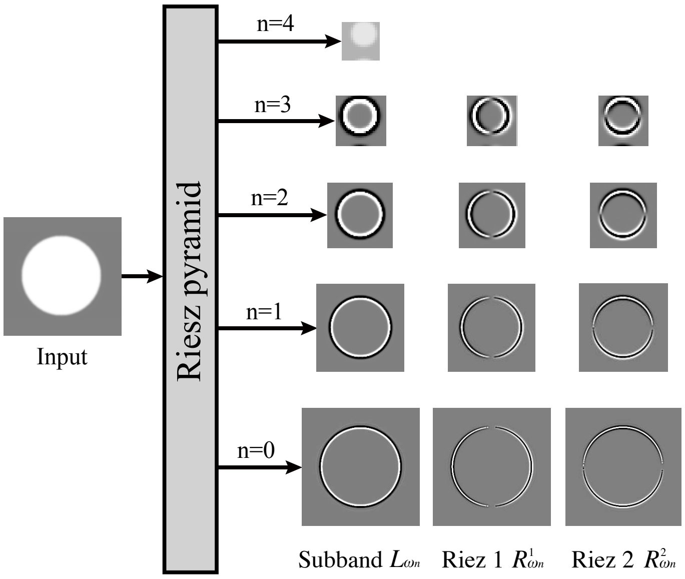


Figure 6.1: Riesz pyramid proposed in [6]. It is built on a set of three images: a subband image signal $L_{\omega_n}(x, y, t)$ and two Riesz-transformed subband image signals $R^1_{\omega_n}(x, y, t)$ and $R^2_{\omega_n}(x, y, t)$

for constructing the Riesz pyramid. Considering $\mathfrak{L}_{\omega_n}(\omega_x, \omega_y, t)$ has most of its subband's energy at the center of spatial angular frequency as $\sqrt{\omega_x^2 + \omega_y^2} = \frac{\pi}{2}$, Wadhawa et al. approximated the Riesz transform by spatial convolution (actually, it is spatial cross-correlation because of non-causality in image signals unlike temporal signals) with three tap finite difference filters $[0.5, 0, -0.5]$ and $[0.5, 0, -0.5]^\top$ as

$$\begin{aligned} h(k) &= \begin{cases} 0.5, & k = -1, \\ 0, & k = 0, \\ -0.5, & k = 1, \end{cases} \\ R^1_{\omega_n}(x, y, t) &\approx \sum_{k=-1}^{1} h(k) \cdot L_{\omega_n}(x + k, y, t), \\ R^2_{\omega_n}(x, y, t) &\approx \sum_{k=-1}^{1} h(k) \cdot L_{\omega_n}(x, y + k, t), \end{aligned} \tag{6.2}$$

because these filters $[0.5, 0, -0.5]$ and $[0.5, 0, -0.5]^\top$ have frequency response respectively as

$$\begin{aligned} -i\sin(\omega_x) &\approx -i\frac{\omega_x}{\sqrt{\omega_x^2+0^2}} = -i\frac{\omega_x}{|\omega_x|}, \\ -i\sin(\omega_y) &\approx -i\frac{\omega_y}{\sqrt{0^2+\omega_y^2}} = -i\frac{\omega_y}{|\omega_y|}, \end{aligned} \tag{6.3}$$

when $\omega_x, \omega_x \approx \frac{\pi}{2}$. Note that this approximation does not have the original 2D filter response of Riesz transform due to the approximation of $\sqrt{\omega_x^2+\omega_y^2} \approx |\omega_x|$ or $|\omega_y|$.

In this Riesz pyramid, we have relations at each (x, y) and t as follows.

$$\begin{aligned} L_{\omega_n}(x,y,t) &= A_{\omega_n}(x,y,t)\cdot\cos\left(\phi_{\omega_n}(x,y,t)\right), \\ R^1_{\omega_n}(x,y,t) &= A_{\omega_n}(x,y,t)\cdot\sin\left(\phi_{\omega_n}(x,y,t)\right)\cdot\cos\left(\theta_{\omega_n}(x,y,t)\right), \\ R^2_{\omega_n}(x,y,t) &= A_{\omega_n}(x,y,t)\cdot\sin\left(\phi_{\omega_n}(x,y,t)\right)\cdot\sin\left(\theta_{\omega_n}(x,y,t)\right), \\ A_{\omega_n}(x,y,t) &= \sqrt{\left(L_{\omega_n}(x,y,t)\right)^2+\left(R^1_{\omega_n}(x,y,t)\right)^2+\left(R^2_{\omega_n}(x,y,t)\right)^2}, \end{aligned} \tag{6.4}$$

where $A_{\omega_n}(x, y, t)$ is a subband amplitude signal, $\phi_{\omega_n}(x, y, t)$ is a subband phase signal, and $\theta_{\omega_n}(x, y, t)$ is the dominant steerable orientation in which the phase signal occurs. From Eq. (6.4), the phase signal $\phi_{\omega_n}(x, y, t)$ can be obtained as

$$\phi_{\omega_n}(x,y,t) = \tan^{-1}\left(\frac{\sqrt{\left(R^1_{\omega_n}(x,y,t)\right)^2+\left(R^2_{\omega_n}(x,y,t)\right)^2}}{L_{\omega_n}(x,y,t)}\right). \tag{6.5}$$

In the complex steerable pyramid used by conventional EVM methods [5, 7, 8, 9, 48, 51], θ has to be fixed by a user in advance, e.g. eight orientations $\theta = \left\{0, \frac{1}{8}\pi, ..., \frac{7}{8}\pi\right\}$, and $\phi_{\omega_n}(x, y, t)$ is parameterized as $\phi_{\omega_n,\theta}(x, y, t)$ with θ. Thus, the complex steerable pyramid is over-complete with respect to θ. In contrast, the Riesz pyramid removes the arbitrariness of the orientations and only detects phase signals $\phi_{\omega_n}(x, y, t)$ with respect to the dominant orientation $\theta_{\omega_n}(x, y, t)$ at every pixel, time frame, and pyramid level. Therefore, the Riesz pyramid lowers the over-completeness and achieves a faster pyramid construction.

This phase signal can be regarded as $S_{\omega_n,\theta}(x,y,t)$ where $\theta = \emptyset$, so we obtain a amplified phase signal $\hat{\phi}_{\omega_n}(x,y,t)$ by using the conventional EVM methods [9, 48, 51] that we explained before this chapter. Finally, we obtain the amplified subband image signal $\hat{L}_{\omega_n}(x,y,t)$ with $\hat{\phi}_{\omega_n}(x,y,t)$ of Eq. (6.5) and Eq. (6.4), and then collapse the amplified subbands pyramid $\{\hat{L}_{\omega_n}(x,y,t) \mid n = 0,\dots,N-1\}$ to reconstruct a magnified image signal $\hat{I}(x,y,t)$ by following the reverse procedure of constructing the non-oriented subbands pyramid. This PbEVMM method with the Riesz pyramid can achieve the faster motion magnification than with the complex steerable pyramid used in conventional EVM methods [5, 9, 48, 51]. However, since the entire image pixel positions must be processed, the Riesz pyramid requires a long computational time in proportion to the video resolution and the number of image time frames.

6.3 Proposed Method

For further accelerating the computational time of the Riesz pyramid [6], we propose a faster PbEVMM method that combines local image processing with the conventional fast PbEVMM method with Riesz pyramid [6] when analyzing subtle phase signals in a video. On the basis of a correlation of phase signals between adjacent pyramid levels as reported in [16, 17], our method constructs fewer image pyramid representations than those constructed in the Riesz pyramid. Therefore, we first consider the correlation of phase signals as shown in Fig. 6.2.

Figure 6.2 shows the local phase signals in the same image areas between adjacent pyramid levels behave similarly as $\phi_{\omega_n}(x,y,t) \approx \lambda \cdot \phi_{\omega_{n+1}}(x,y,t)$ with the pyramid scaling factor λ. Note that large error will occur in this correlation if the pyramid level is too far away. From this consideration, we noticed that we only have to process minimum number of sufficient local image areas at a pyramid level n re-

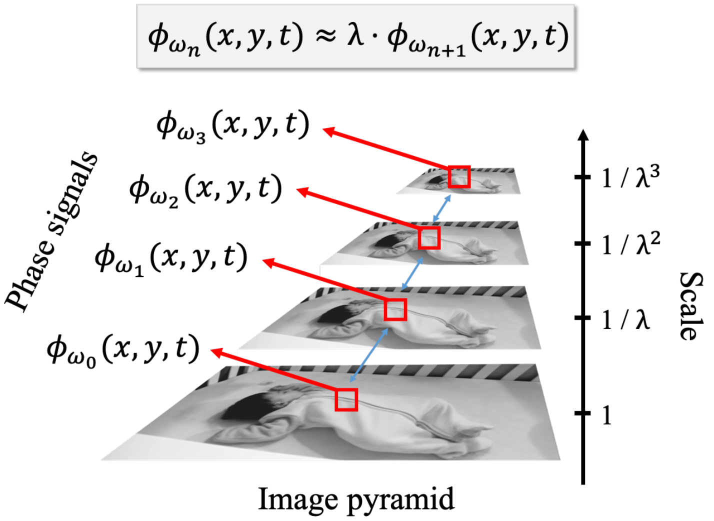


Figure 6.2: Correlation of local phase signals between adjacent pyramid levels. If we observe local phase signals in the same image areas between pyramid levels, phase signals at the upper pyramid level are smaller than those at the lower level with pyramid scaling factor λ as $\phi_{\omega_n}(x,y,t) \approx \lambda \cdot \phi_{\omega_{n+1}}(x,y,t)$. Note that large error will occur in this correlation if the pyramid level is too far away. (the figure is a modified version of [16] in presentation)

lated to strongly magnified image areas at the above $n+1$. Therefore, we propose local Riesz pyramid, which automatically processes the minimum number of sufficient local image areas to quickly produce the motion magnification results (black and red line flow in Fig. 6.3).

Given $\{L_{\omega_n}(x,y,t) \mid n = 0,\ldots,N-1\}$ as the same way of the Riesz method [6], we consider a set of adjacent pyramid levels where the correlation of local phase signals strongly exists: odd- ($n = 2k+1$) and even-numbered ($n = 2k$) with $k = 0,\ldots,\frac{N}{2}-1$. Note that we assume that N is an even-number. In our local Riesz pyramid, we first perform the PbEVMM algorithm with the Riesz pyramid to produce $\hat{L}_{\omega_n}(x,y,t)$ at only odd-numbered $n = 2k+1$, and then calculate a binary image $B_{\omega_{2k+1}}(x, y) \in \{0, 1\}^{H_{2k+1} \times W_{2k+1}}$ as follows.

$$\delta_{2k+1}(x, y) = \sum_{t=1}^{T} \left(\hat{L}_{\omega_{2k+1}}(x, y, t) - L_{\omega_{2k+1}}(x, y, t) \right)^2 , \tag{6.6}$$

$$B_{2k+1}(x, y) = \begin{cases} 1, & \delta_{2k+1}(x, y) > \varepsilon, \\ 0, & \text{otherwise}, \end{cases} \tag{6.7}$$

where ε is a threshold automatically calculated by using Otsu's thresholding method [60] in an implementation of OpenCV [62] to divide image areas at $2k + 1$ into the strongly magnified ones $B_{2k+1}(x, y) = 1$ and not $B_{2k+1}(x, y) = 0$. Note that considering all time frames $t = 1, ..., T$ in Eq. (6.7) is for focusing on cyclic subtle motions from the beginning to the end of an input video rather than those in a short period of time.

After that, we define a set of pixel positions $\mathcal{P}^n$ and divide the positions into $U \times V$ subsets like a grid as

$$\mathcal{P}^n = \{(x, y) \mid 1 \leq y \leq H_n, 1 \leq x \leq W_n\} = \{P^n_{11}, ..., P^n_{UV}\} , \tag{6.8}$$

where P^n_{uv} is defined as

$$P^n_{uv} = \left\{ (x, y) \;\middle|\; 1 + \frac{v-1}{V} W_n \leq x \leq \frac{v}{V} W_n, 1 + \frac{u-1}{U} H_n \leq y \leq \frac{u}{U} H_n, \right\} . \tag{6.9}$$

Note that P^n_{uv} should have 3×3 or more pixel positions to ensure a minimum grid size. We then collect subsets at $2k$ that correspond to the strongly magnified image areas at $2k + 1$ detected by Eq. (6.7) as

$$\mathcal{P}^{2k}_{2k+1} = \left\{ P^{2k}_{uv} \mid \exists (x, y) \in P^{2k+1}_{uv}, B_{2k+1}(x, y) = 1, u = 1, \ldots, U, v = 1, \ldots, V \right\}, \tag{6.10}$$

which means that P^{2k}_{uv} is collected if at least one pixel position $\exists (x, y) \in P^{2k+1}_{uv}$ meets with $B_{2k+1}(x, y) = 1$ at the above pyramid level $2k + 1$.

From Eq. (6.10), we perform the PbEVMM algorithm with the Riesz pyramid in only $(x, y) \in \mathcal{P}^{2k}_{2k+1}$ at $n = 2k$ and then obtain a partially amplified subband image signal $\hat{L}^p_{\omega_n}(x, y, t)$ at $n = 2k$. Finally, we collapse the amplified subbands pyramid $\{\hat{L}_{\omega_n}(x, y, t) \mid n = 2k + 1, k = 0, \ldots, \frac{N}{2} - 1\} \cup \{\hat{L}^p_{\omega_n}(x, y, t) \mid n = 2k, k = 0, \ldots, \frac{N}{2} - 1\}$ to reconstruct a magnified image signal $\hat{I}(x, y, t)$ by following the reverse procedure of constructing the non-oriented subbands pyramid (here, it is Laplacian pyramid).

6.3.1 Generalized Local Riesz Pyramid

Here, we generalize our local Riesz pyramid in terms of how many pyramid levels are the target of our local image processing in Eqs. (6.7) to (6.10). For the generalization, we allow large error of the similarity of local phase signals across pyramid levels as $\phi_{\omega_n}(x, y, t) \approx \lambda \cdot \phi_{\omega_{n+1}}(x, y, t) \approx \lambda^2 \cdot \phi_{\omega_{n+2}}(x, y, t) \cdots \approx \lambda^{N-1} \cdot \phi_{\omega_{n+N-1}}(x, y, t)$ as shown in Fig. 6.2. Then, given a set of $M \in \{M \mid dM = N, d \in \mathbb{N}\}$ pyramid levels with $n = Mk, \ldots, Mk + M - 1$ and $k = 0, \ldots, \frac{N}{M} - 1$, we can generalize Eqs. (6.6), (6.7), and (6.10) as

$$\delta_{Mk+M-1}(x, y) = \sum_{t=1}^{T} \left(\hat{L}_{\omega_{2k+1}}(x, y, t) - L_{\omega_{2k+1}}(x, y, t) \right)^2, \tag{6.11}$$

$$B_{Mk+M-1}(x, y) = \begin{cases} 1, & \delta_{Mk+M-1}(x, y) > \varepsilon, \\ 0, & \text{otherwise}, \end{cases} \tag{6.12}$$

$$\begin{aligned} \mathcal{P}^{Mk+m}_{Mk+M-1} = \Big\{ P^{Mk+m}_{uv} \mid \exists (y, x) \in P^{Mk+M-1}_{uv}, \\ B_{Mk+M-1}(x, y) = 1, \\ u = 1, \ldots, U, v = 1, \ldots, V \Big\}, \end{aligned} \tag{6.13}$$

where

$$m = \begin{cases} M - 2, \ldots, 0, & M \geq 2, \\ 0, & M = 1. \end{cases} \tag{6.14}$$

Note that the generalized local Riesz pyramid is the case of $M \geq 2$, and the Riesz pyramid [6] is the case of $M = 1$ by defining $(x, y) \in \mathcal{P}_k^k = \mathcal{P}^k$.

From Eq. (6.13), we perform the PbEVMM algorithm with the Riesz pyramid in only $(x, y) \in \mathcal{P}_{Mk+M-1}^{Mk+m}$ at $Mk+m\,(m = M-2, \ldots, 0)$ and then obtain a partially magnified subband image signal $\hat{L}^p_{\omega_{Mk+m}}(x, y, t)$ at $Mk + m\ (m = M - 2, \ldots, 0)$. Finally, we collapse the amplified subbands pyramid $\{\hat{L}_{\omega_n}(x, y, t) \mid n = Mk + M - 1, k = 0, \ldots, \frac{N}{2} - 1\} \cup \{\hat{L}^p_{\omega_n}(x, y, t) \mid n = Mk + m, k = 0, \ldots, \frac{N}{2} - 1, m = M - 2, \ldots, 0\}$ to reconstruct a magnified image signal $\hat{I}(x, y, t)$ by following the reverse procedure of constructing the non-oriented subbands pyramid (here, it is Laplacian pyramid).

We expect a computational time to be shorter when we select a bigger M because it constructs very fewer image pyramid representations by the detection procedures of Eqs. (6.11)-(6.14). However, we should keep in mind that the computational cost would be high when M is big if the large local image areas are selected in the first pyramid level $Mk + M - 1$ in the set of pyramid levels M.

6.4 Results

6.4.1 Algorithmic Time Complexity

In this subsection, we analyzed algorithmic time complexity, i.e., how much our proposed local Riesz pyramid reduces the computational time in comparison with the Riesz pyramid [6]. Our local Riesz pyramid and the Riesz pyramid are completely different in terms of analyzing phase signals in whole or local image areas described as Eqs. (6.7)-(6.10). Therefore, we explore the algorithmic time complexity in that process.

Table 6.1 shows algorithmic time complexity for each process: local image processing of Eqs. (6.6)-(6.10), the Riesz transform of Eq. (6.1), or the PbEVMM algo-

Table 6.1: Algorithmic time complexity for each process: local image processing of Eqs. (6.6)-(6.10), the Riesz transform of Eq. (6.1), or the PbEVMM algorithm that contains temporal bandpass filtering of Eq. (3.2) and addition processes of Eq. (3.3). Note that $\left|\mathcal{P}^{2k}_{2k+1}\right|$ is the size of local image areas detected by the local image processing of Eqs. (6.6)-(6.10).

Method	Pyr. level	Local image processing	Riesz transform	PbEVMM algorithm
		Eqs. (6.6)-(6.10)	Eq. (6.1)	Eqs. (3.2)-(3.3)
Riesz [6]	n	–	$\mathcal{O}(H^n W^n T)$	$\mathcal{O}(H^n W^n T \log T)$
Ours	$2k+1$	$\mathcal{O}(H^{2k+1} W^{2k+1} T)$	$\mathcal{O}(H^{2k+1} W^{2k+1} T)$	$\mathcal{O}(H^{2k+1} W^{2k+1} T \log T)$
	$2k$	–	$\mathcal{O}\left(\left\|\mathcal{P}^{2k}_{2k+1}\right\| T\right)$	$\mathcal{O}\left(\left\|\mathcal{P}^{2k}_{2k+1}\right\| T \log T\right)$

rithm that contains temporal bandpass filtering of Eq. (3.2) and addition processes of Eq. (3.3). PbEVMM methods using the Riesz pyramid [6] or our local Riesz pyramid has the longest computational time during the PbEVMM algorithm containing temporal filtering process of Eq. (3.2). By assuming that T is long enough, the algorithmic time complexity strongly depends on the number of image areas $H^n W^n$ and time frame T in the temporal filtering process of Eq. (3.2). For further analysis, we focused on the number of image areas at each time frame t in the temporal filtering process of Eq. (3.2) because it is the critical point for computational time of this algorithm.

The Riesz pyramid [6] processes the entire Laplacian pyramid $\{L_{\omega_n}(x, y, t) \mid n = 0, \ldots, N-1\}$ at t in Eq. (3.2). Thus, its algorithmic time complexity can be defined with respect to the number of image areas of $\{L_{\omega_n}(x, y, t) \mid n = 0, \ldots, N-$

1} as follows.

$$\begin{aligned} g_1 &= \sum_{n=0}^{n=N-1} H^n W^n T \log T \\ &= \sum_{n=0}^{n=N-1} \left(\frac{1}{\lambda^2}\right)^n H^0 W^0 T \log T \\ &= \frac{\lambda^2}{\lambda^2 - 1} \left\{ 1 - \left(\frac{1}{\lambda^2}\right)^N \right\} H^0 W^0 T \log T, \end{aligned} \tag{6.15}$$

where g_M indicates the algorithmic time complexity with respect to the number of image areas in Eq. (3.2). This equation indicates the case of the Riesz pyramid ($M = 1$) [6].

In contrast, our generalized local Riesz pyramid processes the partial Laplacian pyramid $\{\hat{L}^p_{\omega_n}(x, y, t) \mid n = Mk + m, k = 0, \ldots, \frac{N}{2} - 1, m = M - 2, \ldots, 0\}$ at t in Eq. (3.2). By assuming that $|\mathcal{P}^{Mk+m}_{Mk+M-1}| = \frac{1}{q}\left|\mathcal{P}^{Mk+m}\right|$ where $q \in \mathbb{R}_+$, the algorithmic time complexity is described as follows.

$$\begin{aligned} g_M = &\sum_{k=0}^{k=N/M-1} \sum_{m=0}^{m=M-2} \frac{1}{q} H^{Mk+m} W^{Mk+m} T \log T \\ &+ H^{Mk+M-1} W^{Mk+M-1} T \log T \\ = &\sum_{k=0}^{k=N/M-1} \sum_{m=0}^{m=M-2} \frac{1}{q} \left(\frac{1}{\lambda^2}\right)^{Mk+m} H^0 W^0 T \log T \\ &+ \left(\frac{1}{\lambda^2}\right)^{Mk+M-1} H^0 W^0 T \log T \\ = &\left[\frac{\lambda^2}{q\left(\lambda^2 - 1\right)} \left\{ 1 - \left(\frac{1}{\lambda^2}\right)^{M-1} \right\} + \left(\frac{1}{\lambda^2}\right)^{M-1} \right] \\ &\cdot \frac{\lambda^{2M}}{\lambda^{2M} - 1} \left\{ 1 - \left(\frac{1}{\lambda^2}\right)^N \right\} H^0 W^0 T \log T \end{aligned} \tag{6.16}$$

Note that this equation holds for $M \geq 2$. Finally, from Eqs. (6.15) and (6.16), we get the ratio of algorithmic time complexity of the generalized Local Riesz pyramid to the Riesz pyramid [6] as

$$\frac{g_M}{g_1} = \left[\frac{\lambda^2}{q\left(\lambda^2 - 1\right)} \left\{ 1 - \left(\frac{1}{\lambda^2} \right)^{M-1} \right\} + \left(\frac{1}{\lambda^2} \right)^{M-1} \right] \frac{\lambda^{2M}\left(\lambda^2 - 1\right)}{\left(\lambda^{2M} - 1\right)\lambda^2}. \tag{6.17}$$

Fig. 6.4 shows $\frac{g_M}{g_1}$ of Eq. (6.17) with different parameters $\frac{1}{q} = 0, 0.01, \ldots, 1$, $M = 1, 2, 3, 6$, and $\lambda = 2, \frac{4}{3}$. This figure shows that $\frac{g_M}{g_1}$ simply increases linearly in proportion to $\frac{1}{q}$ because local image areas are linearly increasing. In Fig. (6.4) (b), the use of half-octave Gaussian pyramid $\lambda = \frac{4}{3}$ has more computational time than the use of an octave Gaussian pyramid $\lambda = 2$ (Fig. 6.4(a)) because the half-octave Gaussian pyramid requires larger image areas. In the case of $\frac{1}{q} = 0.5$, where the local image areas are half of the original ones, $\frac{g_M}{g_1}$ is near 0.6 at every $M = 2, 3, N$. Therefore, under this condition, the proposed local Riesz pyramid is expected to be about 2x faster than the Riesz pyramid [6]. Moreover, the algorithmic time complexity decreases in proportion to M but converges each value that equals the limit of $\frac{g_M}{g_1}$ as $\frac{1}{q}$ approaches zero. This convergence can be described as

$$\lim_{\frac{1}{q} \to 0} \frac{g_M}{g_1} = \frac{\lambda^2 - 1}{\lambda^{2M} - 1}. \tag{6.18}$$

This equation indicates the best case in our algorithm where no local image areas are magnified except for the pyramid level $Mk + M - 1$.

The above analysis of algorithmic time complexity indicates that the worst case of our method is the same computational complexity as the Riesz method [6] (Fig. 6.4, $\frac{1}{q} = 1$) and the best case is converged with Eq. (6.18). Thus, our method does not completely guarantee that the computational time will be reduced. However, our method usually reduces the computational time because the local image areas to be magnified will be detected by Otsu's method [60] thanks to its simplicity and robustness. Therefore, our EVM method with local Riesz pyramid can achieve faster motion magnification then with the Riesz pyramid [6].

Table 6.2: Parameters for all videos: amplification factor α, target frequency bands between f_1 - f_2, sampling rate f_s.

Video	α	$f_t = [f_1, f_2]$	f_s
baby	25	$[0.5, 1.5]$	30
throat	50	$[100, 120]$	2000
car engine	25	$[0.5, 1.5]$	25
balance	20	$[1.5, 3.0]$	30
drum	20	$[15, 35]$	200
simulation	1∼10	$[9, 11]$	60

6.4.2 Real Videos

To evaluate the effectiveness of our proposed method, which magnifies subtle motions within a short computational time, we conducted experiments on real videos for qualitative evaluation and synthetic ones with ground-truth magnification results for quantitative evaluation. We compared our PbEVMM method using the local Riesz pyramid ($M = 2$) with an PbEVMM method using the Riesz pyramid proposed by Wadhwa et al. [6]. We set the parameters for each experiment as listed in Table 6.2. We performed each method in YIQ color space and divided a set of image areas $\mathcal{P}^n$ into $U \times V = 20 \times 20$ subsets. In all experiments, we specified the ideal bandpass filter as the temporal filter in Eq. (3.2), and pyramid level N as 6. All experiments were implemented by using C++ with OpenCV [62] and ran on a PC with an Intel Core i7-8559U CPU at 2.7 GHz, and 16 GB of RAM.

In Fig. 6.5, our objective was to reveal subtle chest motions caused by the baby's breathing. Comparing our method and the method of [6], both can reveal the subtle chest motions, and thus almost the same video magnification results can be obtained (see right panels in Fig. 6.5). The trend of this qualitative result was also observed in all other experiments (Figs. 6.6, 6.7, 6.8, and 6.9), so our method with the local Riesz pyramid can achieve magnification results that are as good as the method with the Riesz pyramid [6] despite only the local areas being processed.

In Fig. 6.10 and 6.11, we calculated the mean square error (MSE) to check the approximation error of our proposed local Riesz pyramid ($M = 2$) or the generalized one ($M = 3, N$) against the Riesz pyramid [6] as a ground-truth over all image pixel positions, time frames, and color channels with a different value of M. Our proposed method (b) showed low MSE around image areas of baby's chest and the those of the center of drum's membrane, respectively; thus, it can detect the minimum number of sufficient local image areas for revealing principal subtle motions in the input videos. On the other hand, the MSE increased in proportion to M (c,d), which is the case of the generalized local Riesz pyramid (in particular, $M = N$ is the case for which we chose all pyramid levels except for the top $N - 1$ pyramid level). These results indicate that the large error of the phase signals' similarity (Fig. 6.2) in proportion to M directly affected the approximation error between the our proposed and the conventional methods.

Table 6.3 shows the computational time and MSE against the Riesz pyramid [6]. We produced each magnification video result by using our proposed method ($M = 2$), the generalized one ($M = 3, N$), or the method with the Riesz pyramid [6]. This table confirms that our proposed PbEVMM method with the local Riesz pyramid ($M = 2$) requires a shorter computational time in processing an input video than with the Riesz pyramid [6], and has the lowest MSE between $M = 2, 3, N$. Note that, for the baby and the simulation videos in $M = N$, large local image areas were chosen at the first $Nk + N - 1$ pyramid level, thus requiring a long computational time compared with $M = 2$ or 3. Remarkably, our method often achieved almost half the computational time needed to process a video compared with [6]. In our experiments, it is considered that the local image areas are detected as being about

Table 6.3: Comparison of a computational time and MSE against the conventional Riesz pyramid [6]. In all real videos, our proposed method ($M = 2$) required a shorter computational time than the method with the Riesz pyramid [6] and also had MSE lower than $M = 3, N$.

Video	Riesz pyr. [6]	Ours, $M = 2$		$M = 3$		$M = N$	
(H^0, W^0, T)	comp. time (s)	time (s)	MSE	time (s)	MSE	time (s)	MSE
baby $(544, 960, 301)$	31.62	12.79	4.94	6.33	9.69	8.95	13.43
throat $(1144, 720, 300)$	40.56	16.89	3.62	12.98	4.27	12.23	6.31
car engine $(452, 888, 300)$	20.99	13.68	5.73	10.63	11.83	8.42	54.48
balance $(384, 272, 300)$	5.37	3.24	12.94	1.89	28.31	1.17	54.22
drum $(360, 640, 450)$	15.15	9.81	6.70	7.26	13.97	5.82	82.58
simulation $(512, 512, 240)$	9.04	3.37	0.77	2.47	2.16	3.31	8.30

half the size of the original one based on Eq. (6.17).

6.4.3 Controlled Experiments

To evaluate the effectiveness of our proposed method qualitatively, we conducted controlled experiments to assess MSE over all image pixel positions and time frames between a magnified synthetic video by each magnification method (the Riesz pyramid [6], ours $M = 2$, and $M = 3, N$) and the ground-truth (Fig. 6.12). In this experiment, we set the pyramid level N to 6. Fig. 6.12 (top left) shows a

4-second synthetic ball video. The ball had horizontal subtle motions defined as $d = 0.5 \cdot sin(2\pi \frac{f}{f_s} t)$. To obtain a ground-truth magnification video for the synthetic one, we created it while changing d to $5 \cdot d$.

Fig. 6.12 (top center) shows the MSE with a different amplification factor $\alpha = 1, ..., 10$, and the top right plot is an expanded version of the blue rectangle area in the top center plot. In the top right of this figure, there is almost no difference in MSE between our method ($M = 2$) and the method with the Riesz pyramid [6], so this indicates that our proposed method with the local Riesz pyramid can automatically process the minimum number of sufficient local image areas to perform PbEVMM. In contrast, as M increases, the difference in MSE against Riesz pyramid [6] spreads due to mis-detecting local image areas with the large error of the phase change's similarity (Fig. 6.2). Additionally, the effects of the MSE appears in the spatiotemporal slices along the red line (the middle panels) and a green line (the bottom panels) in the input video (the top left plot). All methods can detect subtle ball motions at the left edge of the ball (the bottom panels), but they were ambiguously detected in the cases of $M = 3$ and N at the top edge of the ball (the middle panels). Note that all methods minimize the MSE at the amplification factor $\alpha = 5$, which is consistent with the relationship between the synthetic video and the ground-truth (changing d to $5 \cdot d$). On the other hand, our method outperforms the method with the Riesz pyramid [6] at a high amplification factor $\alpha \geq 6$ because it processes only local image areas and prevents unnecessary magnification outputs.

In Fig. 6.13, we evaluated the effect of an input video that has long time frames on a PbVMM method with the Riesz pyramid [6], our proposed method with the local Riesz pyramid ($M = 2$), and the generalized one ($M = N$). We consider the effect of the time summation process of Eq. (6.6), which is affected by increasing the number of time frames, on a computational time is trivial because the temporal bandpass filtering of Eq. (3.2) is dominant for the computational time in a big O notation manner, see Table 6.1. However, in this experiment, we checked the effect of an input video that has long time frames on our method from an experimental point of view. In this experiment, we evaluated a computational time and MSE against ground-truth with setting the same experimental conditions as the above control experimental condition except for the resolution $(256, 256, T)$ where $T = 240, \ldots, 24000$, $N = 5$, and $\alpha = 5$. Fig. 6.13 left shows that the all PbEVMM methods have linear increase of the computational time in proportion to the number of input time frames. Note that our proposed method's MSE is stable for all input time frames (Fig. 6.13 right). This result indicates that the effect of the number of time frames in an input video on our method is trivial; we can sufficiently ignore the time delay due to the time summation process of Eq. (6.6). Therefore, both the conventional method with the Riesz pyramid [6] and our proposed method with the local Riesz pyramid simply react to the increase or the decrease in the number of input time frames (Fig. 6.13) and pixel positions (Fig. 6.4).

6.5 Discussions and Limitations

We focused on the correlation of phase signals between adjacent pyramid levels and proposed a novel pyramid called local Riesz pyramid that automatically processes the minimum number of sufficient local image areas for PbEVMM. Our method enables to output good video motion magnification results with a shorter computational time than the conventional fast method [6]. It is expected that our method will spark the application of EVM for practical applications where high-speed processing is needed, but there are a number of limitations as follows.

Our proposed method achieves good video motion magnification results equivalent to the conventional method with the Riesz pyramid [6] despite the fact that our method processes only local image areas at several pyramid levels. We consider this is because our method can identify no-motion local image areas and the local image areas where subtle motions exist via Eq. (6.10). However, the boundaries of the local image areas are considered to produce the negative effect of producing discontinuous results. Fortunately, the boundary effect is hardly seen because our local processing is applied only to even-numbered pyramid levels (Figs. 6.10, 6.11 (b)). In contrast, as M increases, the boundary effect clearly appears and leads to high MSE against the Riesz pyramid [6] (Figs. 6.10,6.11 (c, d)). One possible approach to further reducing this boundary effect is weighting the amplification factor α near the boundaries using a Gaussian distribution, but we need to propose a radical way of overcoming this problem in the future.

In our proposed method, local image areas to be processed are estimated by using all image frames as that in Eq. (6.6) because we focused on cyclic subtle motions from the beginning to the end of an input video, rather than those in a short period of time. This suggests that we implicitly assume that subtle motions to be revealed will stay in the same local image areas over all time frames, in other words, objects do not move largely. Therefore, to reveal subtle motions under the presence of large motions within a short computational time, we need to develop a method where the local image areas to be processed are adaptively determined in each image frame without mis-detection of subtle motions.

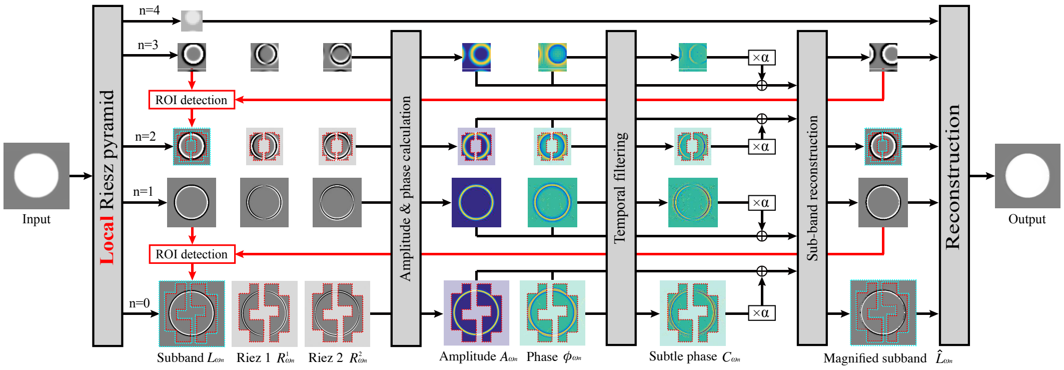


Figure 6.3: Flow chart of PbEVMM method with the Riesz pyramid [6] (black line flow) and with our proposed local Riesz pyramid (black and red line flow). Method with the Riesz pyramid [6] cnostructs a subband image signal $L_{\omega_n}(x, y, t)$ and applies the Riesz transform to all the subband image signal as in Eq. (6.1) and the subsequent processes are performed (black line flow). In contrast, our proposed method first performs all processes only at odd-numbered pyramid levels. Then, strongly magnified local image areas are detected by using ROI detection procedure defined as Eqs. (6.6)-(6.7) and gird-like collection defined as Eq. (6.10) (for details, see Section 6.3). With correlation of local phase signals between adjacent pyramid levels (Fig. 6.2), it is considered that local image areas at the below pyramid (red dot areas) that correspond to the strongly magnified local image areas at the upper pyramid level are also strongly magnified. In contrast, no-motion local image areas are not magnified clearly (cyan dot areas), so it is reasonable to exclude no-motion local image areas in terms of process (transparent areas). Thus, our method processes only local image areas at the below pyramid level and achieves a short computational time for obtaining magnification results.

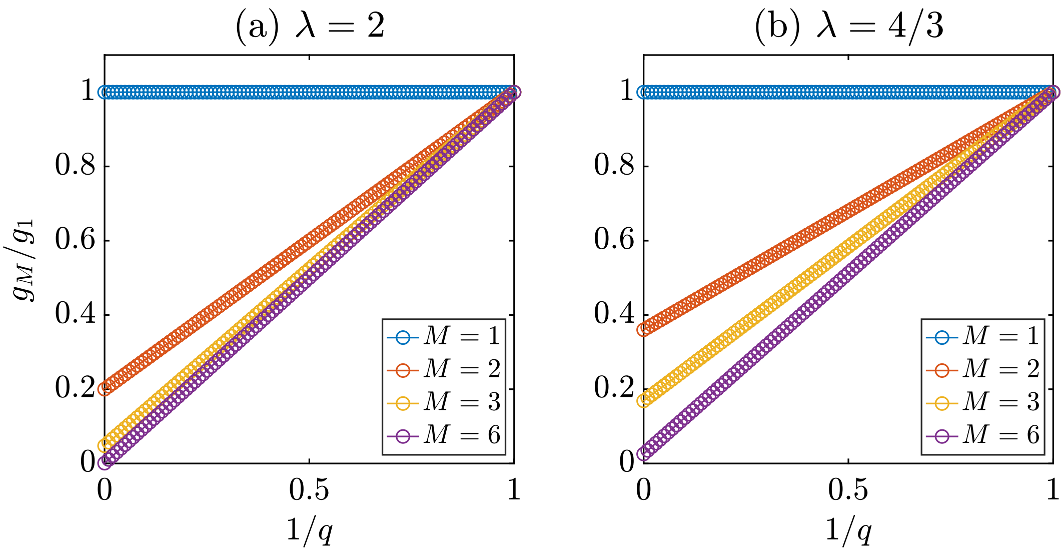


Figure 6.4: The ratio of algorithmic time complexity $\frac{g_M}{g_1}$ of our generalized local Riesz pyramid g_M to the conventional Riesz pyramid g_1 [6] with respect to $\frac{1}{q}$ that determines size of local image areas.

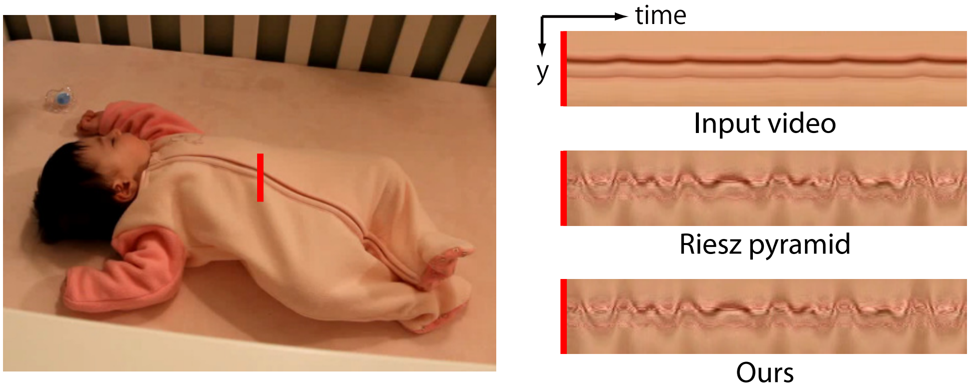


Figure 6.5: The breathing of a baby: visualizing subtle chest motions. We show the spatiotemporal slices (right panels) along a single red line in the left panel. Both a method with the Riesz pyramid [6] and our proposed method with the local Riesz pyramid can reveal the subtle chest motions in the baby (see the right panels).

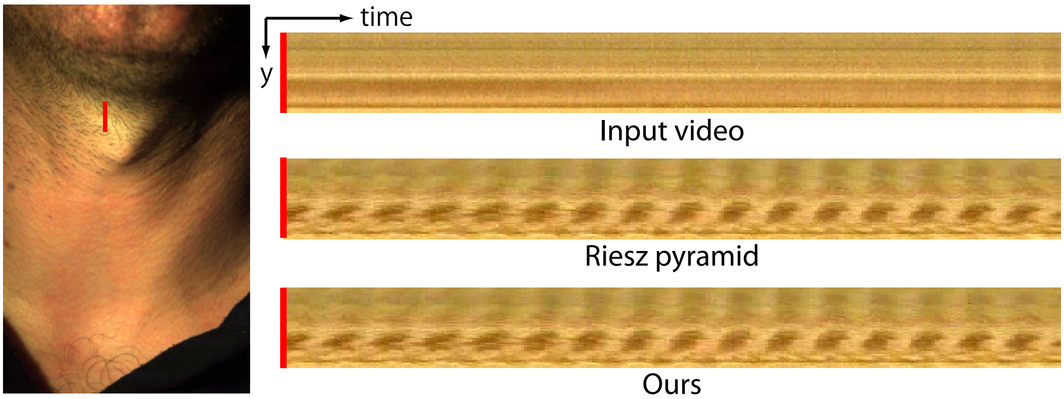


Figure 6.6: Video magnification for revealing subtle skin vibrations of a stationary man who is speaking. We show the spatiotemporal slices (right panels) along a single red line in the left panel. Both a method with the Riesz pyramid [6] and our proposed method with the local Riesz pyramid can reveal the subtle skin vibrations (see the right panels).

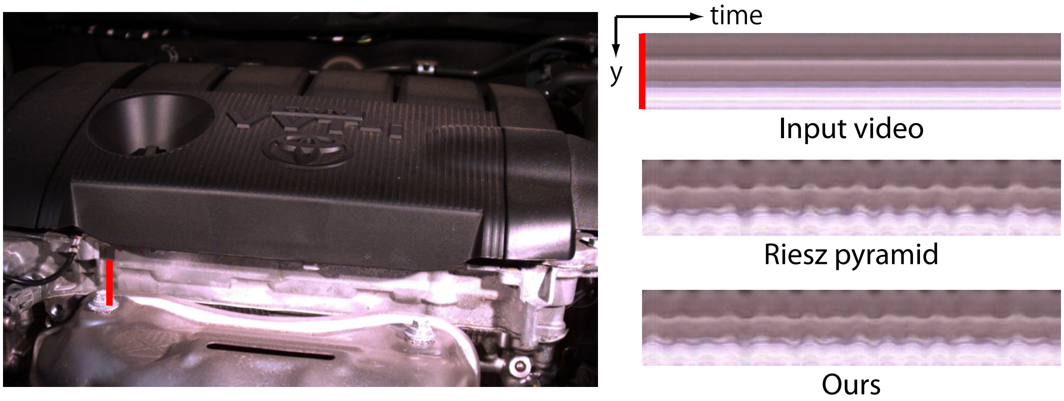


Figure 6.7: A car engine: revealing subtle cyclic vibrations. We show the spatiotemporal slices (right panels) along a single red line in the left panel. Both the method with the Riesz pyramid [6] and our proposed method with the local Riesz pyramid can reveal the subtle cyclic vibrations in the car engine (see the right panels).

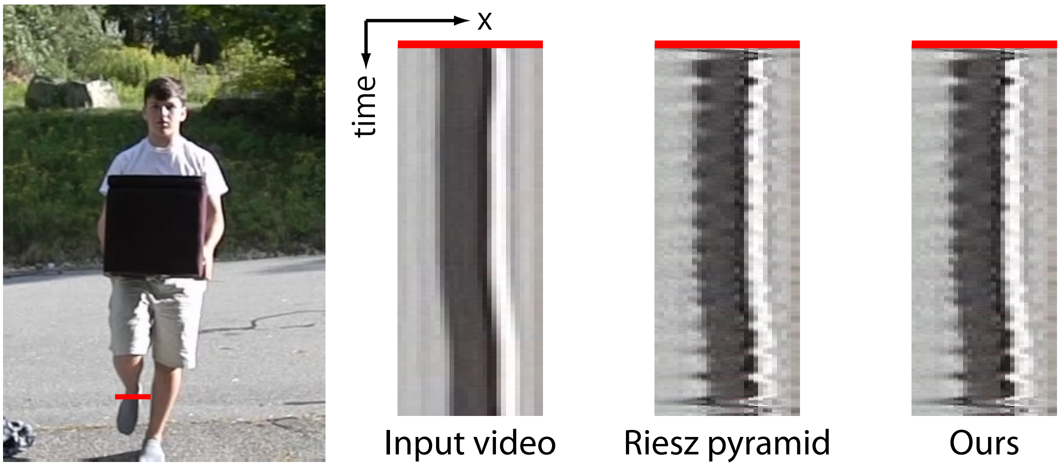


Figure 6.8: A stationary man with a luggage: revealing subtle tremors of a man in balance. We show the spatiotemporal slices (right panels) along a single red line in the left panel. Both the method with the Riesz pyramid [6] and our proposed method with the local Riesz pyramid can reveal the subtle tremors of the man in balance (see the right panels).

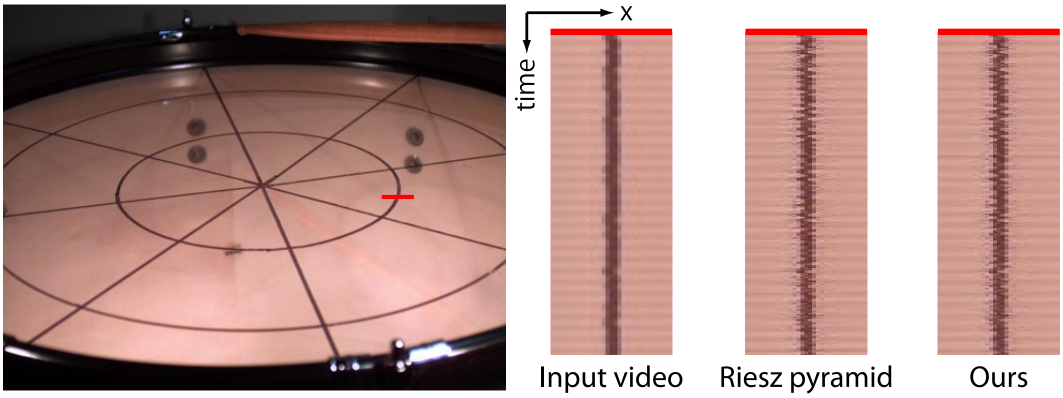


Figure 6.9: Video magnification for revealing subtle membrane vibrations of a drum. We show the spatiotemporal slices (right panels) along a single red line in the left panel. Both the method with the Riesz pyramid [6] and our proposed method with the local Riesz pyramid can reveal the subtle membrane vibrations (see the right panels).

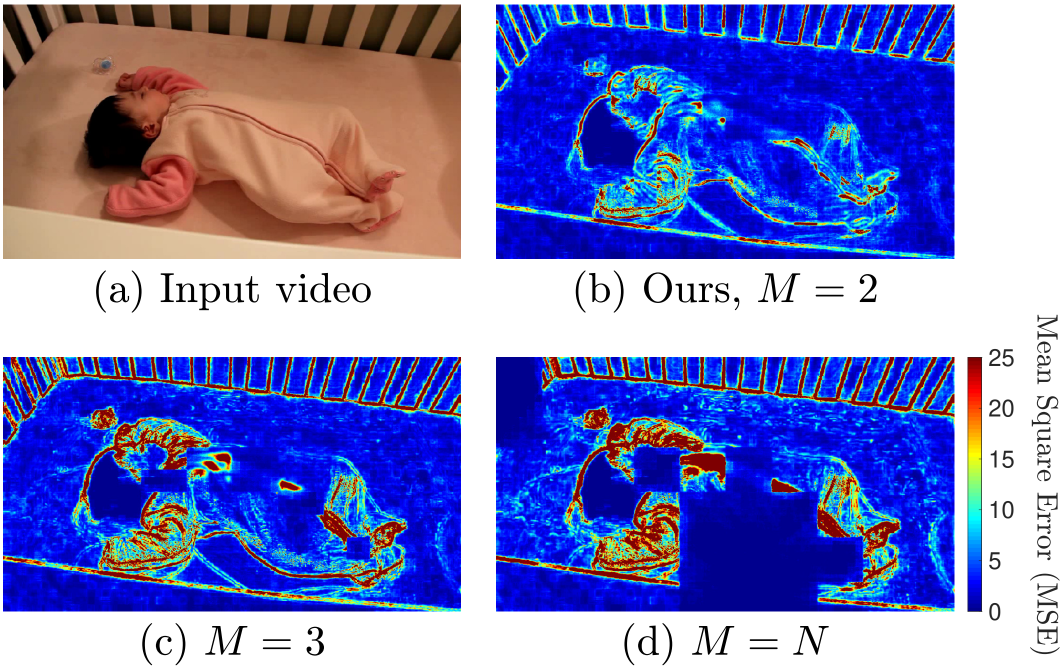


Figure 6.10: Mean square error (MSE) between the proposed local Riesz pyramid and the conventional Riesz pyramid [6] as ground-truth. Our proposed method (b) shows lower MSE around image areas of baby's chest compared with higher M; thus, it can detect the minimum number of sufficient local image areas for revealing principal subtle chest motions.

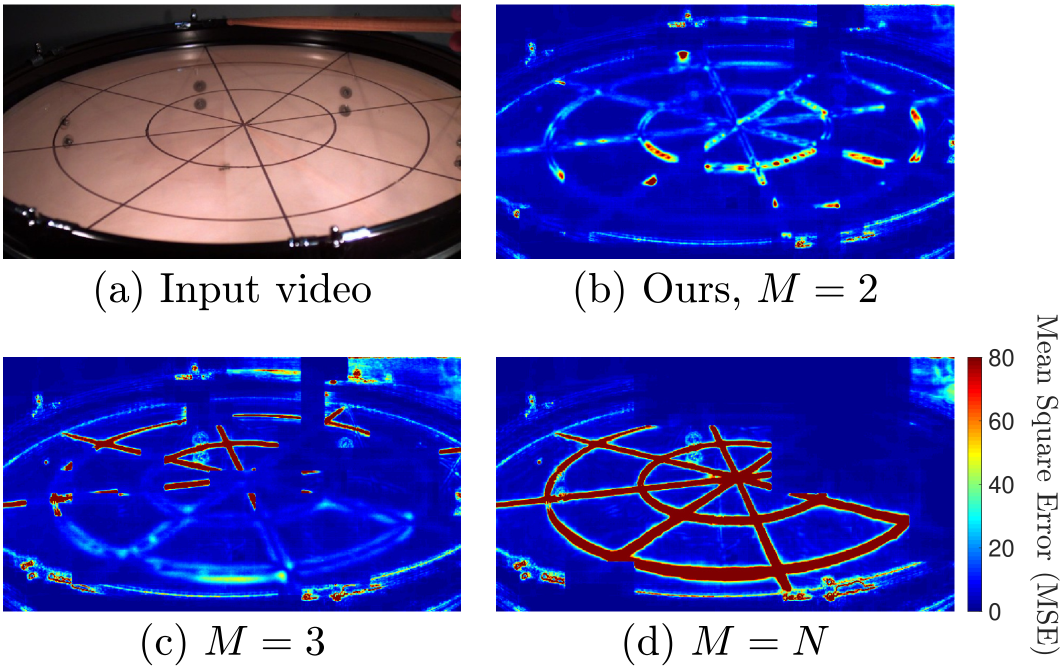


Figure 6.11: Mean square error (MSE) between the proposed local Riesz pyramid and the conventional Riesz pyramid [6] as ground-truth. Our proposed method (b) shows lower MSE around image areas of center of drum's membrane compared with higher M; thus, it can detect the minimum number of sufficient local image areas for revealing principal subtle skin vibrations.

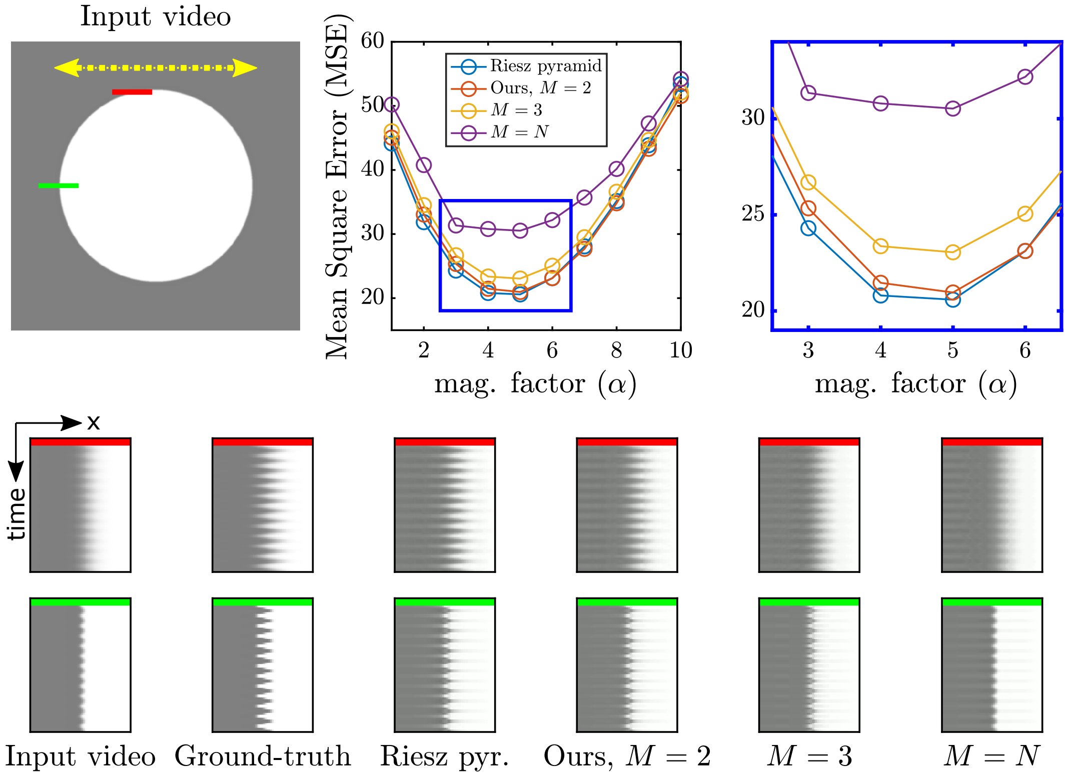


Figure 6.12: Left top: a synthetic ball video. The yellow arrow indicates horizontal subtle motions of the ball defined as $d = 0.5 \cdot sin(2\pi \frac{f}{fs} t)$. Top center and top right plots: MSE with the ground truth for the magnified ball video (smaller MSE is better). Middle and bottom panels: spatiotemporal slices along a single red green line in the top left panel. Our proposed method $M = 2$ has almost same MSE result as the method of [6] at each magnification factor. Note that all methods show the lowest MSE at the magnification factor 5, which is consistent with the controlled experimental condition.

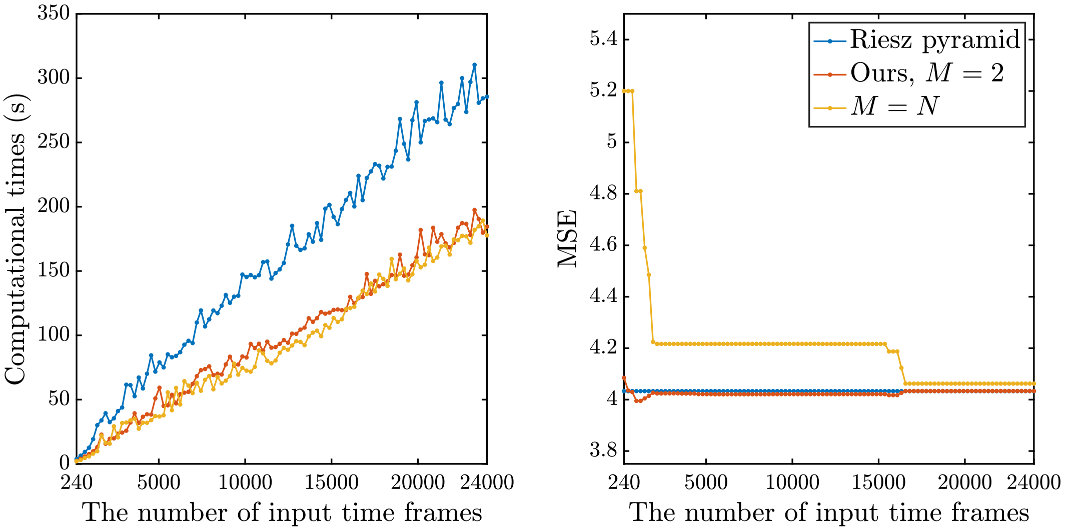


Figure 6.13: Effect of input video that has long time frames on an method with the the Riesz pyramid [6], our proposed method with the local Riesz pyramid ($M = 2$), and the generalized one ($M = N$). A computational time increases linearly in proportion to the number of input time frames (left) with relatively stable MSE (right). Thus, all methods do not have special response to the number of input time frames.

Chapter 7

Conclusion

In this dissertation, we focused on the goal of enhancing the performance of Eulerian video magnification (EVM) for practical applications where only subtle color changes or motions caused by physical/natural phenomena need to be revealed quickly and correctly. To this goal, we facilitated the robust and fast analysis of subtle changes in a video with overcoming the three EVM problems (Problems 1, 2, and 3) as introduced in Section 1. Our methods proposed in this dissertation succeeded in clearly revealing subtle yet important physical/natural phenomena imperceptible to the naked eye even under the practical conditions where large motions of objects exist (Chapter 4), photographic subtle noise in a video exist (Chapter 5), and short computational time is required (Chapter 6). Thus, our methods enable users to easily obtain correct insights and conclusions for practical applications. We verified effectiveness of our methods in the extensive experiments including the various real and simulation videos. Overall, this dissertation makes several important contributions and left future work as described in this chapter.

7.1 Contributions

We state contributions for each chapter as follows.

- **Ignoring Large Motions in Video.** We have proposed an EVM method that ignores large motions of objects and reveals only subtle color changes or motions in a video. While the EVAM method [9] ignores only slow large motions of objects, our method uses jerk, which has been used to evaluate smoothness of time series data in neuroscience and mechanical engineering fields, to make the EVAM method robust even to quick large motions as well as slow large motions. This jerk-aware EVAM (JAEVAM) method enables users to easily reveal subtle color changes or motions in a video even in the presence of large motions of objects because it does not require burdensome limitations such as human manipulations and/or an special camera settings. Moreover, we give a theoretical view of our method, which explains a new theoretical connection between the conventional EVM methods and ours via local Taylor expansions in the temporal domain. This theoretical view enables uses to easily understand how our JAEVAM method is more effective than the conventional EVM methods.

- **Ignoring Photographic Subtle Noise in Video.** We have proposed an EVM method that ignores photographic subtle noise and reveals only meaningful subtle color changes or motions in a video. In developing our method, we designed fractional anisotropy filter (FAF), which evaluates anisotropic diffusion in temporal distribution of subtle color/phase signals, to detect only meaningful subtle color changes or motions. Moreover, we designed a hierarchical edge-aware regularization (HEAR) for refining uncertain subtle motions at flat (texture-less) regions in a video. Our method, in which FAF and HEAR are applied to the JAEVAM method [48], prevents users from obtaining misleading results in which non-meaningful subtle changes caused by photographic subtle noise exist.

- **Accelerating Computational Time.** We proposed an image pyramid, called local Riesz pyramid, that accelerates a computational time of phase-based Eulerian video motion magnification (PbEVMM). Considering the correlation of phase signals between adjacent pyramid levels, the local Riesz pyramid automatically processes the minimum number of sufficient local image areas at several pyramid levels in order to perform PbEVMM within a short computational time. Thus, the PbEVMM method with our local Riesz pyramid enables users to reveal subtle color changes or motions in a video even under practical applications where high-speed processing is required, such as anomaly detection or medical usage.

7.2 Future Work

We state our future works and our long-term view as follows.

- **Ignoring Large Motions in Video.** In Chapter 4, our research is the first attempt of performing EVM without burdensome interventions under the presence of large motions of objects via multiple sptiotemporal filtering approach. However, as described in Chapter 4, our method slightly disturbs to detect subtle color changes or motions in a video and requires multiple steps. Therefore, developing a more simple and robust method based on sptiotemporal filtering approach can be a subject for future work.

- **Ignoring Photographic Subtle Noise in Video.** As we aforementioned in Chapter 5, our proposed FAF assumes non-meaningful subtle changes caused by photographic subtle noise are sampling from Gaussian distribution.

This limits to handle other noise types (e.g., gamma, exponential, uniform, etc. [58]). Moreover, FAF requires the empirical estimation of covariance and the eigen-decomposition for it. Since the empirical estimation is not robust to outliers under the Gaussian assumption, we should consider to use a robust estimation, e.g., a minimum covariance determinant approach [59], instead of the empirical estimation that we used. On the other hand, the eigen-decomposition has slow computational time with respect to the data size of input videos. Thus, a faster algorithm for our method needs to be developed in future work.

- **Accelerating Computational Time.** The local Riesz pyramid that we proposed in Chapter 6 accelerates the computational time of the PbEVMM method. On the other hand, the local Riesz pyramid locally processes an input video with square grid subsets as Eq. (6.8), and thus it causes discontinuous boundaries of PbEVMM. As one possible approach to this issue, we can mitigate the discontinuous boundaries by using 2D Gaussian smoothing. However, we need to propose a radical way of overcoming this issue in the future. Moreover, this method implicitly assumes that subtle motions to be revealed will stay in the same local image areas over all time frames, in other words, objects do not move largely. Thus, this method cannot be easily used in combination with the EVAM [9] method and our JAEVAM method, which are superior to ignore large motions of objects in a video. Therefore, we need to develop a method where the local image areas to be processed are adaptively determined by each time frame.

- **Multimodal Magnification.** Similar to EVM, there are also many research of revealing essential property hidden in some data, such as audio signal and text. Therefore, we are expecting that the combination of those research can improve each other's accuracy and reveal essential property hidden in the input data more clearly. For example, sound-source enhancement is one of the popular research tasks in audio signal processing. This research aims to enhance and reveal a target audio signal from input audio signals by using, e.g., beamforming [63], independent component/vector/low-rank matrix analysis [64, 65, 66, 67, 68], time-frequency masking [69], and deep-learning [70, 71]. Therefore, we consider that EVM can be combined with this research through signal processing or deep learning manner to improve each other's accuracy and enhance/reveal target image or audio signals hidden in the input signal data. Additionally, textual enhancement is a common tool used to facilitate users' attention and/or awareness for the specific purpose of the text, e.g., technical documentation and second language development [72]. This research enhances the appearance of specific words (or sometimes sentences) in a document by, e.g., bold-facing, underlining, capitalizing, italicizing, coloring, using different fonts, and different sizes. Therefore, we consider that the enhanced textural information can be the clue to detect target signals in the EVM algorithm, or, the EVM result can be the clue to identify target words of textual enhancement. On the basis of the above research combination, we, in closing, propose a new research concept called "multimodal magnification". We expect this concept to reveal essential property hidden in the input multimedia data among, e.g., image signal, audio signal, and text information, more clearly by utilizing those data similarity, mutual information, and so on through signal processing or deep learning techniques.

The methods we proposed in this dissertation can be used for various practical applications. For example, it has been said that the EVM methods can potentially be used to analyze structural integrity of buildings, bridges, and car-engines via revealing their subtle vibrations. However, it becomes more practical by our methods in terms of robustness to contamination of large motions of objects (Chapter 4) and that of photographic subtle noise (Chapter 5), or fast computational time (Chapter 6). Moreover, we believe that our methods' robustness to the disturbances in a video has opened a new practical applications of contactless vital-sign monitoring via camera in the wild, e.g., medical field, law enforcement, and disaster relief.

References

[1] Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. Remote plethysmographic imaging using ambient light. *Optics Express*, 16(26):21434–45, 2008.

[2] Ming-Zher Poh, Daniel J. McDuff, and Rosalind W. Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics Express*, 18(10):10762–10774, 2010.

[3] Chen Wang, Thierry Pun, and Guillaume Chanel. A comparative survey of methods for remote heart rate detection from frontal face videos. *Frontiers in Bioengineering and Biotechnology*, 6:33, 2018.

[4] Hao-Yu Wu, Michael Rubinstein, Eugene Shih, John Guttag, Frédo Durand, and William Freeman. Eulerian video magnification for revealing subtle changes in the world. *ACM Transactions on Graphics (TOG)*, 31(4), 2012.

[5] Neal Wadhwa, Michael Rubinstein, Frédo Durand, and William T. Freeman. Phase-based video motion processing. *ACM Transactions on Graphics (TOG)*, 32(4):80:1–80:10, 2013.

[6] Neal Wadhwa, Michael Rubinstein, Frédo Durand, and William T. Freeman. Riesz pyramids for fast phase-based video magnification. In *The IEEE International Conference on Computational Photography (ICCP)*, pages 1–10, 2014.

[7] Mohamed A. Elgharib, Mohamed Hefeeda, Frédo Durand, and William T. Freeman. Video magnification in presence of large motions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4119–4127, 2015.

[8] Julian F.P. Kooij and Jan C. van Gemert. Depth-aware motion magnification. In *The European Conference on Computer Vision (ECCV)*, pages 467–482, 2016.

[9] Yichao Zhang, Silvia L. Pintea, and Jan C. van Gemert. Video acceleration magnification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 502–510, 2017. https://acceleration-magnification.github.io/.

[10] David J Fleet and Allan D Jepson. Computation of component image velocity from local phase information. *International journal of computer vision*, 5(1):77–104, 1990.

[11] Tamar Flash and Neville Hogans. The coordination of arm movements: An experimentally confirmed mathematical model. *Journal of neuroscience*, 5(7):1688–1703, 1985.

[12] Brandon Rohrer, Susan Fasoli, Hermano Igo Krebs, Richard Hughes, Bruce Volpe, Walter R Frontera, Joel Stein, and Neville Hogan. Movement smoothness changes during stroke recovery. *Journal of neuroscience*, 22(18):8297–8304, 2002.

[13] Rieko Osu, Kazuko Ota, Toshiyuki Fujiwara, Yohei Otaka, Mitsuo Kawato, and Meigen Liu. Quantifying the quality of hand movement in stroke patients through three-dimensional curvature. *Journal of NeuroEngineering and Rehabilitation*, 8(1):62, 2011.

[14] Susumu Mori and Jiangyang Zhang. Principles of diffusion tensor imaging and its application to basic neuroscience research. *Neuron*, 51:527–539, 2006.

[15] Andrew L. Alexander, Jee Eun Lee, Mariana Lazar, and Aaron S. Field. Diffusion tensor imaging of the brain. *Neurotherapeutics*, 4(3):316–329, 2007.

[16] Piotr Didyk, Pitchaya Sitthi-Amorn, William Freeman, Frédo Durand, and Wojciech Matusik. Joint view expansion and filtering for automultiscopic 3d displays. *ACM Transactions on Graphics (TOG)*, 32(6):221:1–221:8, 2013.

[17] Simone Meyer, Oliver Wang, Henning Zimmer, Max Grosse, and Alexander Sorkine-Hornung. Phase-based frame interpolation for video. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1410–1418, 2015.

[18] Ce Liu, Antonio Torralba, William T. Freeman, Frédo Durand, and Edward H. Adelson. Motion magnification. *ACM Transactions on Graphics (TOG)*, 24(3):519–526, 2005.

[19] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI)*, volume 2, page 674–679, 1981.

[20] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. "GrabCut": Interactive Foreground Extraction Using Iterated Graph Cuts. *ACM Transactions on Graphics (TOG)*, 23(3):309–314, 2004.

[21] Alexei A. Efros and Thomas K. Leung. Texture synthesis by non-parametric sampling. In *The IEEE International Conference on Computer Vision (ICCV)*, volume 2, pages 1033–1038, 1999.

[22] Jerome Revaud, Philippe Weinzaepfel, Zaid Harchaoui, and Cordelia Schmid. Epicflow: Edge-preserving interpolation of correspondences for optical flow. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1164–1172, 2015.

[23] Till Kroeger, Radu Timofte, Dengxin Dai, and Luc Van Gool. Fast optical flow using dense inverse search. In *The European Conference on Computer Vision (ECCV)*, pages 471–488, 2016.

[24] Aviram Bar-Haim and Lior Wolf. Scopeflow: Dynamic scene scoping for optical flow. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7998–8007, 2020.

[25] Gedas Bertasius and Lorenzo Torresani. Classifying, segmenting, and tracking object instances in video with mask propagation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9739–9748, 2020.

[26] Zili Yi, Qiang Tang, Shekoofeh Azizi, Daesik Jang, and Zhan Xu. Contextual residual aggregation for ultra high-resolution image inpainting. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7508–7517, 2020.

[27] Tae-Hyun Oh, Ronnachai Jaroensri, Changil Kim, Mohamed Elgharib, Frédo Durand, William T Freeman, and Wojciech Matusik. Learning-based video motion magnification. In *The European Conference on Computer Vision (ECCV)*, pages 663–679, 2018.

[28] Dorkenwald Michael, Buchler Uta, and Ommer Bjorn. Unsupervised magnification of posture deviations across subjects. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8256–8266, 2020.

[29] Berthold K.P. Horn and Brian G. Schunck. Determining optical flow. *Artificial Intelligence*, 17(1):185 – 203, 1981.

[30] L. Liu, L. Lu, J. Luo, J. Zhang, and X. Chen. Enhanced eulerian video magnification. In *2014 7th International Congress on Image and Signal Processing*, pages 50–54, 2014.

[31] Eero P Simoncelli and William T Freeman. The steerable pyramid: A flexible architecture for multi-scale derivative computation. In *The IEEE International Conference on Image Processing (ICIP)*, volume 3, pages 444–447, 1995.

[32] William T Freeman, Edward H Adelson, et al. The design and use of steerable filters. *IEEE Transactions on Pattern analysis and machine intelligence*, 13(9):891–906, 1991.

[33] Javier Portilla and Eero P Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International journal of computer vision*, 40(1):49–70, 2000.

[34] Jue Wang, Steven M. Drucker, Maneesh Agrawala, and Michael F. Cohen. The cartoon animation filter. *ACM Transactions on Graphics (TOG)*, 25(3):1169–1173, 2006.

[35] Jiamin Bai, Aseem Agarwala, Maneesh Agrawala, and Ravi Ramamoorthi. Selectively de-animating video. *ACM Transactions on Graphics (TOG)*, 31(4):66, 2012.

[36] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 23(11):1222–1239, 2001.

[37] Vivek Kwatra, Arno Schödl, Irfan Essa, Greg Turk, and Aaron Bobick. Graphcut textures: Image and video synthesis using graph cuts. *ACM Transactions on Graphics (TOG)*, 22(3):277–286, 2003.

[38] Martin Fuchs, Tongbo Chen, Oliver Wang, Ramesh Raskar, Hans-Peter Seidel, and Hendrik P. A. Lensch. Real-time temporal shaping of high-speed video streams. *Computers & Graphics*, 34(5):575–584, 2010.

[39] Abe Davis, Michael Rubinstein, Neal Wadhwa, Gautham J. Mysore, Frédo Durand, and William T. Freeman. The visual microphone: Passive recovery of sound from video. *ACM Transactions on Graphics (TOG)*, 33(4):79:1–79:10, 2014.

[40] Abe Davis, Katherine L. Bouman, Justin G. Chen, Michael Rubinstein, Oral Büyüköztürk, Fredo Durand, and William T. Freeman. Visual vibrometry: Estimating material properties from small motions in video. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 39(4):732–745, 2017.

[41] Neal Wadhwa, Justin G. Chen, Jonathan B. Sellon, Donglai Wei, Michael Rubinstein, Roozbeh Ghaffari, Dennis M. Freeman, Oral Büyüköztürk, Pai Wang, Sijie Sun, Sung Hoon Kang, Katia Bertoldi, Frédo Durand, and William T. Freeman. Motion microscopy for visualizing and quantifying small motions. *Proceedings of the National Academy of Sciences (PNAS)*, 114(44):11639–11644, 2017.

[42] Tianfan Xue, Jiajun Wu, Zhoutong Zhang, Chengkai Zhang, Joshua B Tenenbaum, and William T Freeman. Seeing Tree Structure from Vibration. In *European Conference on Computer Vision (ECCV)*, pages 762–779, 2018.

[43] Sergey Tulyakov, Xavier Alameda-Pineda, Elisa Ricci, Lijun Yin, Jeffrey F. Cohn, and Nicu Sebe. Self-adaptive matrix completion for heart rate estima-

tion from face videos under realistic conditions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2396–2404, 2016.

[44] Guha Balakrishnan, Fredo Durand, and John Guttag. Detecting pulse from head motions in video. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3430–3437, 2013.

[45] Aurelio Piazzi and Antonio Visioli. Global minimum-jerk trajectory planning of robot manipulators. *IEEE Transactions on Industrial Electronics*, 47(1):140–149, 2000.

[46] Tony Lindeberg. *Scale-Space Theory in Computer Vision*, volume 256. Kluwer Academic Publishers, 1994.

[47] Krystian Mikolajczyk and Cordelia Schmid. Indexing based on scale invariant interest points. In *The IEEE International Conference on Computer Vision (ICCV)*, volume 1, pages 525–531, 2001.

[48] Shoichiro Takeda, Kazuki Okami, Dan Mikami, Megumi Isogai, and Hideaki Kimata. Jerk-aware video acceleration magnification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1769–1777, 2018.

[49] Ankur Datta, Mubarak Shah, and N Da Vitoria Lobo. Person-on-person violence detection in video data. In *The International Conference on Pattern Recognition (ICPR)*, volume 1, pages 433–438, 2002.

[50] Yinlin Hu, Rui Song, and Yunsong Li. Efficient coarse-to-fine patchmatch for large displacement optical flow. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5704–5712, 2016.

[51] Shoichiro Takeda, Yasunori Akagi, Kazuki Okami, Megumi Isogai, and Hideaki Kimata. Video magnification in the wild using fractional anisotropy in temporal distribution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1614–1622, 2019.

[52] Manisha Verma and Shanmuganathan Raman. Edge-aware spatial filtering-based motion magnification. In *Proceedings of 2nd International Conference on Computer Vision & Image Processing*, pages 117–128, 2018.

[53] Xiu Wu, Xuezhi Yang, Jing Jin, and Zhao Yang. Pca-based magnification method for revealing small signals in video. *Signal, Image and Video Processing*, 12:1293–1299, 2018.

[54] Sylvain Paris, Samuel W. Hasinoff, and Jan Kautz. Local laplacian filters: Edge-aware image processing with a laplacian pyramid. *ACM Transactions on Graphics (TOG)*, 30(4):68:1–68:12, 2011.

[55] Storyblocks.com. www.videoblocks.com, 2019.

[56] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, , and Andrea Vedaldi. Describing textures in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3606–3613, 2014.

[57] Ce Liu, William T. Freeman, Richard Szeliski, and Sing Bing Kang. Noise estimation from a single image. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 901–908, 2006.

[58] Ajay Boyat and Brijendra Joshi. A review paper: Noise models in digital image processing. *Signal & Image Processing : An International Journal (SIPIJ)*, 6(2):63–75, 2015.

[59] Peter J. Rousseeuw. Least median of squares regression. *Journal of the American Statistical Association*, 79(388):871–880, 1984.

[60] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979.

[61] Keith Langley and Stephen J. Anderson. The riesz transform and simultaneous representations of phase, energy and orientation in spatial vision. *Vision Research*, 50(10):1748–1765, 2010.

[62] Gary Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[63] Jack Capon. High-resolution frequency-wavenumber spectrum analysis. *Proceedings of the IEEE*, 57(8):1408–1418, 1969.

[64] Hiroshi Sawada, Ryo Mukai, Shoko Araki, and Shoji Makino. Frequency-domain blind source separation. In *Speech Enhancement*, chapter 13, pages 299–327. Springer Berlin Heidelberg, 2005.

[65] Taesu Kim, Hagai T. Attias, Soo-Young Lee, and Te-Won Lee. Blind source separation exploiting higher-order frequency dependencies. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(1):70–79, 2007.

[66] Daichi Kitamura, Nobutaka Ono, Hiroshi Sawada, Hirokazu Kameoka, and Hiroshi Saruwatari. Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(9):1626–1641, 2016.

[67] Hiroshi Sawada, Nobutaka Ono, Hirokazu Kameoka, Daichi Kitamura, and Hiroshi Saruwatari. A review of blind source separation methods: two con-

verging routes to ilrma originating from ica and nmf. *APSIPA Transactions on Signal and Information Processing*, 8(e12):1–14, 2019.

[68] Shoichiro Takeda, Kenta Niwa, and Shinya Shimizu. Differentiable max-directivity beamforming normalization for independent vector analysis. In *2020 28th European Signal Processing Conference (EUSIPCO)*, pages 296–300, 2021.

[69] Yariv Ephraim and David Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(6):1109–1121, 1984.

[70] Yuma Koizumi, Kenta Niwa, Yusuke Hioka, Kazunori Kobayashi, and Hitoshi Ohmuro. Informative acoustic feature selection to maximize mutual information for collecting target sources. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(4):768–779, 2017.

[71] DeLiang Wang and Jitong Chen. Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10):1702–1726, 2018.

[72] Shawn Loewen and Solène Inceoglu. The effectiveness of visual input enhancement on the noticing and l2 development of the spanish past tense. *Studies in Second Language Learning and Teaching*, 6(1):89–110, 2016.