Biomedical Signal Processing and Control 34 (2017) 25–35

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Synthesizing the motion of the vocal folds using optical flow based techniques☆

Gustavo Andrade-Miranda[a,*], Nathalie Henrich Bernardoni[b,c], Juan I. Godino-Llorente[a]

[a] Center for Biomedical Technology, Univ. Politécnica de Madrid, Campus de Montegancedo, Crta. M40 km, 38, 28223 Madrid, Spain
[b] Univ. Grenoble Alpes, GIPSA-Lab, F-38000 Grenoble, France
[c] CNRS, GIPSA-Lab, F-38000 Grenoble, France

ARTICLE INFO

Article history:
Received 14 October 2016
Received in revised form 1 December 2016
Accepted 6 January 2017

Keywords:
Glottal dynamics
High-speed videoendoscopy
Motion field
Optical flow
Facilitative playbacks
Motion synthesis

ABSTRACT

Different playbacks have been proposed to synthesize the dynamical information of the vocal folds. However most of them rely on the delineation of the glottal gap using segmentation techniques which is a complex task and usually requires a manual supervision. In order to solve this issue, three new playbacks based on the optical flow computation are presented. Two of them, called Optical Flow Glottovibrogram and Glottal Optical Flow Waveform, analyze the global dynamics; and the remaining one, called Optical Flow Kymogram, analyzes the local dynamics. The reliability of the proposed playbacks is evaluated by comparison with traditional representations, showing a great correlation in shape with the traditional playbacks, and allowing the identification of the most important instants of time, such as closed-states and maximal opening. In addition, they provide complementary information to the common spatio-temporal representations, although the new playbacks are lightly blurred.

1. Introduction

The high-speed videoendoscopy (HSV) has revolutionized laryngeal imaging, allowing us to better understand the glottal dynamics during the phonation process. The HSV technique is capable to acquire the true intra-cycle vibratory behavior which permit the study of cycle-to-cycle glottal variations. HSV let characterize laryngeal tissue dynamics and vocal folds vibratory features, which are not possible to assess (visualize) using common videoendoscopic and stroboscopic techniques [3–5].

HSV records thousands of frames per second, which makes impossible the manual analysis of such amount of information. Therefore, it is needed the use of image processing techniques to synthesize the time-varying data into a few static images.

The literature reports some proposals to represent in a more simple way the HSV information. These representations improve the quantification accuracy, facilitate the visual perception, and increase the reliability of visual rating while preserving the most relevant characteristics of glottal vibratory patterns. These representations are known as *facilitative playbacks* [6]. The most widespread and successful playbacks used either by clinicians or researchers are: Digital Kymograms (DKG) [7], Mucosal Wave Kymogram (MKG) [6], Glottal Area Waveform (GAW) [8], Phonovibrogram (PVG) [9], and Glottovibrogram (GVG) [10]. Depending on the way they assess glottal dynamics, they can be grouped in local- or global-dynamics playbacks.

Local-dynamics playbacks analyze the vocal folds behavior along one single line that is computed on a line perpendicular to the main glottal axis. DKG is the most extended method in this category and it has been successfully applied to demonstrate the change of glottal dynamics in case of damaged tissues, such as lesions, scars, discoloration of the vocal folds and voice disorders [11,12].

On the other hand, global-dynamics playbacks analyze the vocal folds behavior along the whole glottal length, being GAW, PVG and GVG the most wide-spread methods. These three playbacks are focused on vocal folds edge motion by means of glottal segmentation algorithms. For instance, GAW uses the glottal segmentation to compute a glottal area function along time from which several parameters can be estimated [13]. Contrariwise PVG and GVG playbacks are 2D representations of vocal folds vibratory patterns as a function of time, for which glottal-edge movements along the anterior–posterior axis are summarized into a time-varying image

☆ A preliminary version of this work has been reported in INTERSPEECH 2015 and MAVEBA 2015 [1,2].

* Corresponding author.
E-mail addresses: gxandrade@ics.upm.es (G. Andrade-Miranda), Nathalie.Henrich@gipsa-lab.fr (N. Henrich Bernardoni), ignacio.godino@upm.es (J.I. Godino-Llorente).

http://dx.doi.org/10.1016/j.bspc.2017.01.002
1746-8094/

line. In comparison to GVG, PVG allows to distinguish left- and right-fold movements, and is thus more sensitive to the accuracy of glottal main-axis [10]. PVG has been used to classify functional voice disorders [14], vibratory patterns [15], and to discriminate early stages of malignant and precancerous vocal folds lesions [16].

Despite of the usefulness of these playbacks, some drawbacks restrict their applicability. For instance, they are based on glottal-area segmentation, which is a complex task and requires supervision. Moreover, the motion analysis is limited only to the points belonging to the glottal contour. Thus, new methods for data visualization are needed, which would integrate time dynamics, such as velocity or acceleration, and spatial dynamics, such as vocal fold edges and body motion.

In this context, Optical Flow (OF) techniques [17] emerge as an alternative candidate to synthesize vocal folds motion in consecutive frames by creating a motion field in which each pixel represents a vector displacement. OF has been used in a variety of situations to analyze the dynamic properties of tissues or cellular objects, the deformation of organs [18] or cells [19], motion estimation of cardiac ultrasound images [20] and tracking colonoscopy videos [21], among others.

In this paper, a novel approach based on OF computing is proposed for visualizing glottal dynamics of HSV sequences in a compact manner. Firstly, the general principles of OF computation are explained (Section 2.1) discussing its applicability to the Laryngeal HSV problem (Section 2.2). Later, three proposed facilitative playbacks are described to reduce the spatio-temporal dimensionality, focusing on local dynamics (Section 3.3.1), on global dynamics (Section 3.3.2), and on glottal velocity (Section 3.3.3). In order to assess the reliability of the innovative playbacks, the traditional ones based on glottal segmentation are used as a baseline (Section 3.5). The results obtained show that the new playbacks may complement, or even replace, those based on glottal segmentation, providing a visual assessment of both glottal edges and mucosal wave (MW) (Section 4).

The paper is organized as follows. Section 2 details the principles of OF based image processing. Section 3 presents the three playbacks and describes the database of HSV sequences used in the study. Section 4 evaluates the proposed OF playbacks with regard to commonly-used ones. Conclusions and perspectives are given in Section 5.

2. Principles of optical-flow computation

2.1. General principles

Motion analysis aims at understanding and interpreting the dynamical behavior of moving objects. A low-level motion characterization is the estimation of a displacement vector for each pixel in the image, creating a dense motion field that is called Optical Flow (the term OF and motion field are used indistinctly along the paper).

OF estimation has been used for the last 35 years since the seminal works of Horn–Schunck and Lucas–Kanade [22,23], and many innovative methods have been proposed to solve its computation [24]. However, to date, there is no unique method to characterize at minimal computational cost all the possible motion scenarios, including those with disturbing phenomena such as lighting changes, reflection effects, modifications of objects properties, motion discontinuities, or large displacements.

The definition of the OF is originated from a physiological description of the images formed on the retina, which determine that the image is formed due to the change of structured light caused by a relative motion between the eyeball and the scene. In the field of computer vision, Horn–Schunck defined OF in [22] as "*the apparent motion of brightness patterns observed when a camera is moving relative to the objects being imaged*".

Let us denote an image sequence as $I(\mathbf{x}, t)$, where $\mathbf{x} = (x, y) \in \mathbb{R}^2$ represents the position of the pixels and $t \in [0, N]$ is the sampled time interval of the sequence. Hence, the intensity of a given pixel $\mathbf{x}_{ij} = (x_i, y_j)$, $i = \{1, 2, \ldots, n\}$ and $j = \{1, 2, \ldots, m\}$ for an instant t, is denoted as $I(\mathbf{x}_{ij}, t)$. Then, the basic OF assumption is that at a given pixel $\mathbf{x}_{ij}$, at time t, the intensity $I(\mathbf{x}_{ij}, t)$ would remain constant during a short interval of time Δt, the so-called brightness constancy constraint (BCC) or data term (Eq. (1)).

$$I(\mathbf{x}_{ij}, t) = I(\mathbf{x}_{ij} + \vec{\mathbf{w}}(\mathbf{x}_{ij}, t), t + \Delta t) \quad \forall \quad \mathbf{x}_{i,j} \tag{1}$$

where $\vec{\mathbf{w}}(\mathbf{x}_{ij}, t) = (u(\mathbf{x}_{ij}, t), v(\mathbf{x}_{ij}, t))$ is the vector displacement of $\mathbf{x}_{ij}$ in a time Δt. The vector displacement $\vec{\mathbf{w}}(\mathbf{x}_{ij}, t)$ has two components: one in the x-axis direction ($u(\mathbf{x}_{ij}, t)$) and other in the y-axis direction ($v(\mathbf{x}_{ij}, t)$). Therefore, the total motion field at time t is defined as (Eq. (2))

$$\mathcal{W}(\mathbf{x}, t) = \begin{pmatrix} \vec{\mathbf{w}}(\mathbf{x}_{11}, t) & \vec{\mathbf{w}}(\mathbf{x}_{12}, t) & \cdots & \vec{\mathbf{w}}(\mathbf{x}_{1n}, t) \\ \vec{\mathbf{w}}(\mathbf{x}_{21}, t) & \vec{\mathbf{w}}(\mathbf{x}_{22}, t) & \cdots & \vec{\mathbf{w}}(\mathbf{x}_{2n}, t) \\ \vdots & \vdots & \ddots & \vdots \\ \vec{\mathbf{w}}(\mathbf{x}_{m1}, t) & \vec{\mathbf{w}}(\mathbf{x}_{m2}, t) & \cdots & \vec{\mathbf{w}}(\mathbf{x}_{mn}, t) \end{pmatrix} \tag{2}$$

and its components can be defined at the same way respectively by Eq. (3) and (4)

$$U(\mathbf{x}, t) = \begin{pmatrix} u(\mathbf{x}_{11}, t) & u(\mathbf{x}_{12}, t) & \cdots & u(\mathbf{x}_{1n}, t) \\ u(\mathbf{x}_{21}, t) & u(\mathbf{x}_{22}, t) & \cdots & u(\mathbf{x}_{2n}, t) \\ \vdots & \vdots & \ddots & \vdots \\ u(\mathbf{x}_{m1}, t) & u(\mathbf{x}_{m2}, t) & \cdots & u(\mathbf{x}_{mn}, t) \end{pmatrix} \tag{3}$$

$$V(\mathbf{x}, t) = \begin{pmatrix} v(\mathbf{x}_{11}, t) & v(\mathbf{x}_{12}, t) & \cdots & v(\mathbf{x}_{1n}, t) \\ v(\mathbf{x}_{21}, t) & v(\mathbf{x}_{22}, t) & \cdots & v(\mathbf{x}_{2n}, t) \\ \vdots & \vdots & \ddots & \vdots \\ v(\mathbf{x}_{m1}, t) & v(\mathbf{x}_{m2}, t) & \cdots & v(\mathbf{x}_{mn}, t) \end{pmatrix} \tag{4}$$

The BCC provides only one equation to recover the two unknown components of $\mathcal{W}(\mathbf{x}, t)$. Therefore, it is necessary to introduce an additional constraint encoding a priori information of $\mathcal{W}(\mathbf{x}, t)$. Such information comes from the spatial coherency imposed by either local or global constraints (regularization term [25]).

In practice, the BCC assumption is an imperfect photometric expression of the real physical motion in the scene that can not be applied in case of changes in the illumination sources of the scene, shadows, noise in the acquisition process, specular reflections or large and complex deformation. Therefore, several matching costs (also called penalty functions) have been explored to overcome the drawback of the BCC, in particular its sensitivity to noise and illumination changes. For a detailed review about OF techniques, their formulation, regularization and optimization methodology, refer to [25–27].

2.2. Optical flow in laryngeal HSV

Current analysis techniques of the vocal folds motion based on HSV require glottal-edge detection and tracking. Advantageously, OF computation tracks unidentified objects based only on its motion which mean that no segmentation is needed.