Information Sciences 302 (2015) 108–121

Contents lists available at ScienceDirect

Information Sciences

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

Discovery and organization of multi-camera user-generated videos of the same event

Sophia Bano *,[1], Andrea Cavallaro

Centre for Intelligent Sensing, Queen Mary University of London, E1 4NS London, UK

ARTICLE INFO

Article history:
Received 2 February 2013
Received in revised form 7 July 2014
Accepted 7 August 2014
Available online 13 August 2014

Keywords:
Multi-camera synchronization
User-generated video
Audio chroma feature
Video identification
Event clustering
Audio degradation

ABSTRACT

We propose a framework for the automatic grouping and alignment of unedited multi-camera User-Generated Videos (UGVs) within a database. The proposed framework analyzes the sound in order to match and cluster UGVs that capture the same spatio-temporal event and estimate their relative time-shift to temporally align them. We design a descriptor derived from the pairwise matching of audio chroma features of UGVs. The descriptor facilitates the definition of a classification threshold for automatic query-by-example event identification. We evaluate the proposed identification and synchronization framework on a database of 263 multi-camera recordings of 48 real-world events and compare it with state-of-the-art methods. Experimental results show the effectiveness of the proposed approach in the presence of various audio degradations.

1. Introduction

With the increasing availability of smartphones, more people capture videos of their experience of attending events such as concerts, sporting competitions and public rallies. Social media sites then act as a distribution channel to share these experiences by giving access to these *unorganized* and *unsynchronized* User-Generated Videos (UGVs). This trend has invoked a new research direction involving search and organization of multimedia data of the same event [2,40]. We define an *event* as a continuous action captured simultaneously by multiple user-devices from different positions located in proximity with each other.

By identifying videos belonging to a specific event, powerful event browsing can be enabled, which in turn can improve web search tools. However, it is non-trivial to automatically identify UGVs of the same event. In fact traditional metadata-based methods for event retrieval [21,33] may not always be effective because metadata associated with uploaded videos may lack consistent and objective tagging, or correct timestamps [15,23]. Moreover, UGVs are not synchronized, and automatic synchronization is difficult due to the presence of various audio and visual degradations. We are interested in using the audio signal for identifying and synchronizing UGVs. Synchronization of UGVs using audio features is generally based on onsets (starting point of an audio instant) or fingerprints (compact content-based audio signatures) [43,28]. In order for a method to be successful, audio degradations and noise have to be taken into account.

* Corresponding author.
E-mail addresses: sophia.bano@eecs.qmul.ac.uk (S. Bano), andrea.cavallaro@eecs.qmul.ac.uk (A. Cavallaro).

[1] S. Bano was supported by the Erasmus Mundus Joint Doctorate in Interactive and Cognitive Environments, which is funded by the Education, Audiovisual and Culture Executive Agency (FPA n. 2010-0012). This work was supported in part by the UK Engineering and Physical Sciences Research Council (EPSRC), under grant EP/K007491/1.

http://dx.doi.org/10.1016/j.ins.2014.08.026
0020-0255/© 2014 Elsevier Inc. All rights reserved.

We categorize audio degradations into two groups, namely, local and global degradations. *Local degradations* are caused by recording device settings, channel noise, surrounding noise and reverberations. *Global degradations* are common to some or all recording devices (e.g. a crowd cheering, a whistle blowing during a specific event or a public rally) and may help during the synchronization process.

In this paper, we propose an automatic query-by-example event identification and synchronization framework using audio chroma features. Although the recording of a specific event captured by multiple devices might differ in loudness or sound intensity due to the varying quality of recording devices, the distance of the device from the sound source and surrounding noise, the pitch of the recorded sound will remain constant [10]. For this reason, we use chroma as an audio feature [18], as it gives the distribution of energy along different pitch classes. The novelty of this work also lies in the design of a descriptor from match and non-match histograms that facilitates the definition of an automatic classification threshold for event identification and clustering. We show the robustness of the proposed synchronization approach compared to alternative methods over various audio degradations.

The paper is organized as follows. In Section 2, we present the related work. In Section 3, we define and formulate the video identification and synchronization problem. In Section 4, we present an overview of the proposed framework. In Section 5, we describe our proposed event identification framework, which is followed by time-shift estimation and cluster membership validation in Section 6. In Section 7, we describe our dataset of UGVs, assess our method and compare the method with the existing state of the art. Finally, Section 8 concludes the paper.

2. Related work

In this section, we discuss the state of the art for content identification for videos, music and generic sounds, and synchronization for multi-camera videos.

Video identification aims at identifying videos that match with a query, for example, to filter unauthorized distribution of copyrighted videos [22,31,32,35,41,44]. Extending these approaches to UGVs is not trivial because there might not exist the same visual evidence between pairs of UGVs due to variations in the field of view, changing and poor lighting conditions, and visual quality. A related topic is content identification in music for tagging, play-listing and taste profiling [4,8,9,12]. Methods include those used for Shazam and TrackID [29,34,46], which are based on the fingerprinting method by Wang et al. [47] for audio identification.

Event identification using audio features has been addressed in [28,11,5], which use landmark-based audio fingerprinting [47], where the landmarks are the onsets of local frequency peaks and are identified from the short-time Fourier transform. Kennedy and Naaman [28] presented an approach for the synchronization and organization of a collection of concert recordings, in which the classification threshold is computed based on the mean and standard deviation of the matches. Cotton and Ellis [11] used matching pursuit to obtain a prominent representation of audio events and tested their event identification approach on a public speech dataset. Both approaches [28,11] use hash value similarity maximization for matching pairs of recordings. A similar approach is presented by Bryan et al. [5] for event identification and synchronization. This method uses landmark cross-correlation for matching and a fixed classification threshold to cluster a speech dataset of 180 professional recordings and 23 user-generated recordings of concerts. However, a fixed classification threshold [5,28] can be applied only if the dataset under analysis is small.

Existing methods for multi-camera UGV synchronization involve extraction and matching of features such as audio fingerprints [43,28,5,6], audio onsets [43,3], audio feature-based classification [42] and audio-visual events [7], where an audio-visual event was defined to be a simultaneous change in the audio and video streams which are well localized in time. Also, Kammerl et al. [27] proposed graph-based methods for temporal synchronization built by analyzing consistency in pairwise cross-correlation of three audio features, namely, spectral flatness, zero crossing and signal energy. The audio fingerprinting method of Haitsma and Kalker [20] is exploited by Shrestha et al. [42,43]: a 32-bit sub-fingerprint (binary) is generated based on spectrum-temporal analysis of the audio in an overlapping window. Two fingerprint-blocks of 256 consecutive sub-fingerprints are considered to be matching if the number of bit errors (BER) is smaller than a threshold [20]. The landmark-based fingerprinting approach by Wang [47] is used by Kennedy and Naaman [28] and Bryan et al. [5] for the synchronization of collections of concert recordings. However, fingerprinting might become sensitive to reverberations [43] and strong local degradations [36] (see Section 5.3).

In comparison to audio fingerprints, onset-based methods [43] are more sensitive to *audio degradations* as they reflect only positive changes in energy and false positive onsets can be generated by channel and background noise. Casanovas and Cavallaro [7] presented an audio-visual method for multi-camera synchronization in which an audio event is detected using audio onsets [43] followed by visual event detection by analyzing the local variation of pixel intensities within a predefined space–time blocks of a detected audio event. A space–time block is considered to be active if its local variation is greater than a threshold, and an audio-visual event is detected when several active blocks are in close proximity. This method is sensitive to *audio degradations*, in the same way as the onset based method [43] is and is dependent on camera motion and near or far fields of view.

An audio feature classification method for multi-camera synchronization is presented in [42], which is based on low-level signal properties, mel-frequency cepstral coefficients (MFCC), psychoacoustic features (roughness, loudness, sharpness), and temporal envelope fluctuations model. Quadratic discriminant analysis [37] is performed to estimate the probabilities of