Multimed Tools Appl (2014) 70:119–158
DOI 10.1007/s11042-012-1085-1

Multimodal extraction of events and of information about the recording activity in user generated videos

Francesco Cricri · Kostadin Dabov · Igor D. D. Curcio · Sujeet Mate · Moncef Gabbouj

Published online: 5 May 2012

Abstract In this work we propose methods that exploit context sensor data modalities for the task of detecting interesting events and extracting high-level contextual information about the recording activity in user generated videos. Indeed, most camera-enabled electronic devices contain various auxiliary sensors such as accelerometers, compasses, GPS receivers, etc. Data captured by these sensors during the media acquisition have already been used to limit camera degradations such as shake and also to provide some basic tagging information such as the location. However, exploiting the sensor-recordings modality for subsequent higher-level information extraction such as interesting events has been a subject of rather limited research, further constrained to specialized acquisition setups. In this work, we show how these sensor modalities allow inferring information (camera movements, content degradations) about each individual video recording. In addition, we consider a multi-camera scenario, where multiple user generated recordings of a common scene (e.g., music concerts) are available. For this kind of scenarios we jointly analyze these multiple video recordings and their associated sensor modalities in order to extract higher-level semantics of the recorded media: based on the orientation of cameras we identify the region of interest of the recorded scene, by exploiting correlation in the motion of different cameras we detect generic interesting events and estimate their relative position. Furthermore, by

F. Cricri (✉) · K. Dabov · M. Gabbouj
Tampere University of Technology, Tampere, Finland
e-mail: francesco.cricri@tut.fi

K. Dabov
e-mail: kostadin.dabov@gmail.com

M. Gabbouj
e-mail: moncef.gabbouj@tut.fi

I. D. D. Curcio · S. Mate
Nokia Research Center, Tampere, Finland

I. D. D. Curcio
e-mail: igor.curcio@nokia.com

S. Mate
e-mail: sujeet.mate@nokia.com

analyzing also the audio content captured by multiple users we detect more specific interesting events. We show that the proposed multimodal analysis methods perform well on various recordings obtained in real live music performances.

Keywords Multimodal · Indexing · Sensor · Video · Multi-camera · Multi-user · Context · Event

1 Introduction

In the past decade there has been an enormous growth in both the amount of user generated multimedia content (images, video and audio) and in the means for its sharing (over the Internet). This was enabled by rapid advances in the multimedia recording capabilities of mobile devices and by the growth of social networking services. It is thereafter becoming increasingly important to have richer semantic information about the media content in order to enable its effective storage, retrieval, and usage. A typical retrieval use case is searching within a video for particular events or objects which are considered to be interesting or salient. An example of a salient event in a video of a concert is the end of a song performed by the musicians. In order to effectively retrieve such events or objects, indexing the multimedia content is a necessary preliminary step. The indexing process can be performed either manually by a human (by a process known as *manual annotation*) or automatically. Automatic techniques are preferable when the amount of media items to annotate is large. In both manual and automatic cases a preliminary step is required before indexing the media content: the analysis of the content itself in order to *understand the semantics* that are present therein.

Automatic analysis of multimedia data is a very challenging task which has been addressed for many years by researchers from a wide range of scientific fields such as computer science, signal processing, computer vision, pattern recognition, artificial intelligence, etc. The main approach to tackle the multimedia understanding problem has focused on analyzing the video or audio content data, sometimes in a multimodal fashion (as we will discuss in the next section dedicated to the prior art). However, nowadays a large portion of camera-enabled electronic devices (such as modern smartphones) embed various *auxiliary sensors* such as accelerometers, gyroscopes, magnetometers (compasses), Global Positioning System (GPS) receivers, light sensors, proximity sensors, etc., which provide rich *contextual* data. By capturing auxiliary sensor data while the device is recording a video it is possible to obtain additional information about the recorded scene [13]. For example, these sensors are often used during the media acquisition to provide some basic tagging information such as the location used in geo-tagging. As proposed also in [12], by inferring the context of the recording process (for example, how the users move their cameras) it would be possible to reduce the distance between low-level and higher level analysis (known as the "semantic gap"). Furthermore, the results of the sensor data analysis can be used in combination with the results obtained from audio- and video-content analysis.

In this paper we propose a set of methods which exploit auxiliary sensor data for detecting interesting events and inferring information about the recorded activity in user generated videos from individual or multiple cameras. Therefore the contribution of this work is twofold. First, we develop methods for multimodal analysis of each *individual* video, in order to detect:

- *Camera movements*

  - Camera panning.
  - Camera tilting.
- *Video-content degradations*
  - Camera shake.
  - Wrong camera orientation.

Second, we develop multimodal analysis methods applied jointly to videos of the same scene recorded by *multiple* users (for example at the same public happening, such as at a live music show), in order to obtain information about:

- *Angular Region of Interest* of the recorded public happening. We refer to it as the "ROI".
- *Generic interesting events* from co-occurring camera motion in multiple videos:
  - Correlated camera panning.
  - Correlated camera tilting.
- *Relative positions* of those cameras that perform correlated movements with respect to a common object of interest.

In addition, we provide examples of how the detected generic interesting events can be used for discovering the occurrence of more specific interesting events. This is accomplished by combining the auxiliary sensor data analysis with the more traditional multimedia content analysis.

The information extracted by the proposed multimodal analysis methods can then be utilized by several applications or services dealing with multimedia content, in particular those which may benefit from additional knowledge about it. Examples of such applications are:

- *Automatic video retrieval*—Multimedia archives would benefit from the availability of additional information about videos. For example, it would be possible to retrieve a video recorded during a certain camera movement, or a set of videos recorded during a correlated camera panning (which may contain an interesting event, for the reasons that we will explain in Section 5).
- *Automatic video summarization* techniques usually build on the availability of information about salient events which were recorded by the camera(s) [19]. Salient events are the most interesting portions of the video—which summarize the video itself. In this way, a video summary of a public happening consists of only those video segments capturing the salient events.
- *Automatic multi-camera video mash-up generation* is another potential application which makes use of detected interesting events or other information about the video content recorded by multiple cameras. A multi-camera video mash-up is a sequence of video segments captured by different cameras which are aligned in time in order to follow the original timeline. The ultimate goal of methods that automatically generate such video mash-ups is to obtain a result which is similar to what a human (for example a director) could produce. As for the creation of video summaries, this would require knowledge of high-level content information; for example, knowledge about the presence of an interesting event captured by a certain camera would be used for switching view to that camera when the event is taking place. By considering the video content quality, segments of low quality (e.g., shaky segments) could be discarded. Also the knowledge of the mutual distance between different cameras can be exploited in the context of multi-camera video mash-up generation.

An illustration of the extraction and utilization of multimodal information from multiple capturing devices is provided in Fig. 1.

2 Prior art

A significant amount of work has already been done for analyzing video or audio content in order to extract low- and/or high-level descriptors and concepts. A set of relatively low-level descriptors for multimedia indexing and retrieval are made available as part of the MPEG-7 content description standard [20]. As already discussed in the introductory section, in this paper we mainly focus on the analysis of videos for detecting the presence of camera movements, video degradations and interesting events, for estimating the relative distance between cameras and for identifying the region of interest. In the following we briefly describe prior works which address each of these problems, by also highlighting their issues and how our methods attempt to solve them or how their approaches are different from ours. As our paper is based on multimodal analysis, we will describe prior works utilizing the same kind of approach. An interesting review on the start-of-the-art of multimodal analysis is presented in [30]. In the review one of the conclusions that the authors come up with is that "multimodal analysis is the future". Furthermore, whenever possible, we specify the purpose for which the described analysis technique is thought to be used by the authors, i.e., for multimedia indexing and retrieval, for automatic video summarization, or for automatic multi-camera video mash-up production.

Detecting the *camera motion* performed by individual cameras by means of video content analysis has been studied extensively in the last twenty years. Early examples of such works

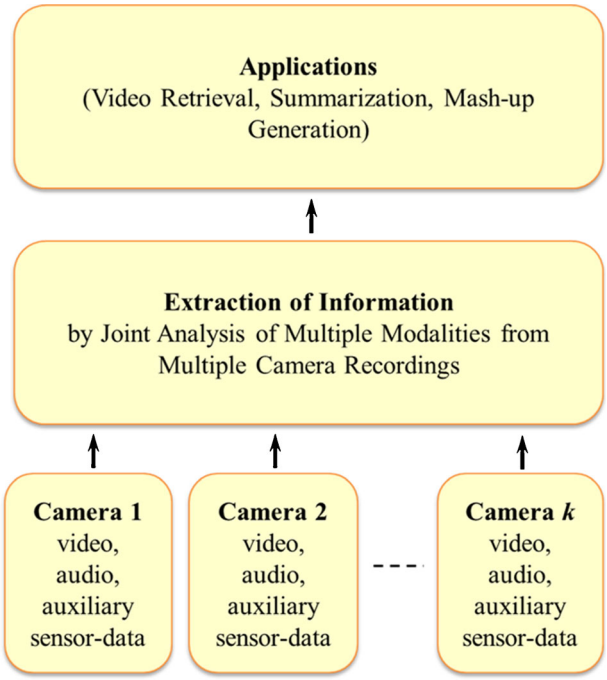

Fig. 1 Overview of the proposed analysis which exploits multiple cameras and multiple modalities

are described in [34] and [16]. A more recent contribution which is of particular interest to our paper, as it considers user generated videos, is the one proposed by Abdollahian et al. [1]. The authors detect the camera movements for the purpose of automatically generating video summaries. In particular, the camera motion is regarded as indicator of the camera person's interests in the scene, and is analyzed by means of content analysis for identifying key-frames and regions of interest within a key-frame. In [10] and [24] camera motion is detected for the purpose of automatically generating video summaries. In [10] fast camera motion is detected as it can cause blurry images. In [24] camera movements are detected for determining scene boundaries. These methods are all based on video-content analysis. This implies two things: they are computationally demanding and their performance is impaired by the presence of moving objects in the scene and by motion blur in the video frames. Instead, the camera motion detection methods that we propose in this paper do not analyze the effects that motion has on the recorded visual data but directly measure the camera movements by using motion sensors.

Video quality estimation plays an important role in all the three purposes of multimedia analysis that we are considering in this paper. In particular, in multimedia indexing and retrieval applications it would be useful to know which video segments have low quality so that they are not retrieved as a result of a query (if not otherwise specified). Furthermore, video summaries and multi-camera mash-ups should not contain any low-quality segments either. Shrestha et al. [28] proposed an automatic video mash-up system that takes into account the video quality. In particular, they use a quality measure [29] which is based on the blurriness, blockiness, shakiness and brightness of the video recordings. All of these four measures are determined from the video content; in particular, the shakiness depends on the motion between consecutive frames; the brightness depends on the pixel intensities within a frame, and the blurriness [22] depends on the average extent of edges in the frame. In [21] the authors propose a novel method for detecting the 3D camera-shake trajectory from the captured image. In [32] home videos are considered. Features based on three motion properties, namely speed, direction and acceleration, are extracted from video segments. Subsequently, such video segments are classified as blurred, shaky, inconsistent or stable using Support Vector Machines. In our method we also analyze the direction of the camera motion and the acceleration, but we obtain them directly from the accelerometer and not from the image content. Instead, the works described in [21, 28] and [32] are based on content analysis. In the context of photography, data captured by motion sensors is exploited for shake analysis in [27], wherein the authors target the detection of non-shaky temporal segments during which allowing the camera to capture a photo. In [8] the authors analyze sensor data in real-time for video stabilization in security surveillance systems. However they do not discuss the details of the actual analysis being performed. In our approach instead we consider user generated videos captured by mobile phones. We analyze streams of recorded sensor readings, representing contextual information of such videos, to detect camera shake throughout each recorded video (and classify this shake according to its strength). In this way, we mark those video segments which have been detected as shaky, for example in order to exclude them in a multi-camera video production or in the results of a query for the case of multimedia retrieval applications. Our method works in a different way than [8] and we will give more details in Section 4.2.1.

Regarding the *estimation of the region of interest* in videos captured by multiple cameras, the prior art has mainly considered visual sensor networks for surveillance or for environment monitoring. Thus, as for our knowledge, no prior works have dealt with estimating the region in which people are interested the most. For surveillance/monitoring types of scenarios it is important to determine the "area of coverage" of the deployed cameras.

Examples of such prior works are given in [18] and in [7]. In [18] the authors describe a method that maximizes the area of coverage of a randomly deployed directional sensor network. In [7] mobile sensors are used in order to reduce the number of required sensors and to increase the robustness against dynamic network conditions. The authors consider as the area of coverage the cell into which a sensor enters from a neighboring cell. These methods are related to what we propose in this paper, but they attempt to achieve a different goal. In fact our region of interest estimation method tries to understand where the angular focus of main attention is in a public happening, without actively controlling the sensors (as instead it is done in [18]). Also, as opposed to [7], we do not aim to reduce the number of recording cameras and our region of interest depends on the directions towards which the cameras are pointed.

Detecting interesting/salient events in videos has been a research topic for long time now, and it is still an open problem as it requires knowledge about the high-level semantics contained in a video. In the following we firstly discuss some of the existing content-based multimodal event detection methods (i.e., those combining the analysis of video, audio and eventually text) for the purpose of *multimedia retrieval*. In [4], events are extracted from news videos in a multimodal unsupervised fashion. This method combines information from audio, visual appearance, faces, and mid-level semantic concepts by applying coherence rules. Some of the considered events are interview, dialog, introduction, commercial, anchor, reporter. In [26] the authors study the joint modeling of text and image components of multimedia content, by exploiting correlations between the two components. The method is applied for cross-modal retrieval, i.e., for retrieving text documents in response to query images and vice-versa. It is shown that the cross-modal model outperforms existing uni-modal image retrieval systems. As in [26], in one of our event detection methods we also exploit correlations between different modalities, namely between motion sensor data and audio-content data. However, the works described in [4] and [26] analyze visual data (apart from other types of data) which is usually computationally expensive. The analysis methods that we propose are instead light-weight, as they mainly consider auxiliary sensor data sampled at low rates, as we will describe in the next sections of this paper. Therefore our methods are very suitable for being employed in large-scale video analysis frameworks.

Event detection is not restricted only to content searches made by users (i.e., to perform video retrieval) but it plays a key role also in automatic *video summarization* applications. In [25] a generic architecture for automated event detection in broadcast sport videos by means of multimodal analysis was proposed. Low-level, mid-level and high-level features were extracted from the audio and visual content (mainly from color and motion information) for detecting events. Some of the high-level features detected were “Audience”, “Field”, “Goal”, “Close-up of player”, “Tennis point” and “Soccer goal”. For detecting an event, the system looks for a certain temporal pattern of high-level features which could be either co-occurring or in temporal sequence.

Automatic or semi-automatic *multi-camera video mash-up* generation is another notable application of high-level events detection. In [15], a system for organization of user-contributed content from live music shows is presented. The authors consider as interesting events those time-intervals for which a higher number of captured videos is available. However, this method does not further analyze each extracted time-interval for detecting more specific interesting events happening within it. Instead we propose to detect such higher-level events by exploiting multiple modalities captured by multiple recording devices. Another related work is presented in [14] and it targets group-to-group videoconferencing. Therein, events (such as the presence of a dialogue) are defined by Ontology Web Language models, and state changes in the resulting video stream are triggered by reasoning

on the event ontology. View selection based on detection of events in multi-camera systems is performed also in [23], where the authors have modeled events specific to an office environment. Low-level features are manually annotated and given to a Bayesian network (BN) for learning the event models. The obtained BN model is then used for the recognition of events by analyzing the features extracted from all cameras. View selection considers the camera which provides the optimal view in terms of recognition probability of given event. Examples of detected events are: "Cleaning", "Conversation", "Meeting", etc. The prior works on multi-camera video mash-up described in [14] and [23] consider "clean" and controlled recording environments (e.g., laboratory environments), where the cameras are still and are capturing a non-noisy audio scene. Instead, we analyze user generated videos which are typically corrupted by unintentional camera motion and where the recording environment is often not known a priori or predictable (e.g., in the case of recording an unanticipated event). A method for detecting spatio-temporal events of interest in single and multi-camera systems is described in [31], in which low-level events (denoted as *primitive* events) are used as building blocks and combined by several available operators (*AND*, *OR*, *SEQUENCE*, *REPEAT-UNTIL*, and *WHILE*) for detecting semantically higher level events (denoted as *composite* events). The events are not predefined and hard-coded; instead they are specified by the users in the form of composite events. Examples of primitive events are the followings: moving objects in a region of interest, directional motion of objects, abandoned objects, object removal, trip wire, and camera blinded/moved. The event detection methods presented in [14, 23] and [31] analyze video and/or audio content data. This can represent a bottleneck, as the analysis would become computationally too complex when the number of videos or cameras is too high. In [3], interesting events are detected when users behave in a similar way in terms of view similarity, group rotation, and acoustic-ambience fluctuation. In addition to detecting interesting events by content analysis, the authors also mention the use of compass to detect what they call group rotation, indicated by a large number of people rotating towards the event. They use a special setup where mobile phones are attached to the clothes of the camera users and are used to record during a happening. Instead in our work we rely solely on user generated content without any restrictions on the camera use.

Regarding prior work on estimating the distance or the position of cameras, no prior work has been done on exploiting the correlated motion of cameras capturing the same (moving) object and relying only on auxiliary sensor data. However, one interesting related method is presented in [2], in which the relative position of non-overlapping cameras is estimated by recovering the missing trajectory points in the unobserved areas. For this, the algorithm uses Kalman filtering and linear regression, but the authors do not explain how to obtain the trajectory data from the audio-visual content (which is postponed to the future work). In [11] the relative position between a device and a stationary object of interest is estimated by fusing vision and inertial sensors (accelerometer and gyro). Our approach does not need any specific landmark for estimating the mutual distance. Also, it is totally independent of the video or audio content, as it analyzes only motion sensor data.

Unlike some of the discussed prior works, the methods for analyzing videos that we propose in this paper consider videos captured by users in a non-controlled environment (i.e., not in a laboratory, but in a real usage scenario), using camera-enabled mobile phones without any restriction on how to handle them. Furthermore, we mainly analyze contextual data (from auxiliary sensors) captured simultaneously with the video recording, thus avoiding performing computationally expensive processing of the visual data. The topic of analyzing contextual data for the purpose of inferring information about media content has received limited attention by the research community. In [6], the authors study the use of geo-location, image content and compass direction for annotating images, for the purpose of

image retrieval. The use of compass (group rotation) and accelerometer (for shake detection) was mentioned in [3]. In particular, accelerometer variance is used to detect if an interesting shot is stable but the authors do not elaborate any further on this topic.

The rest of the paper is organized as follows. In Section 3 we introduce the auxiliary sensors that we use in this work, background information about the estimation of camera motion, and some of the video degradations commonly experienced in user generated video. Section 4 describes the methods that we propose for performing multimedia indexing by individually analyzing sensor data from each camera. In Section 5 we describe our methods for extracting information (i.e., detecting events and estimating relative distances and ROI) by jointly analyzing multimodal data captured by multiple cameras. Experimental evaluation and related discussion are given in Section 6. Finally, Section 7 contains a general discussion on the developed methods, on future directions of our research, and our concluding remarks.

3 Preliminaries

In this section we provide background information that is the basis for the work we describe in the subsequent sections. In particular, we introduce the auxiliary sensors that we consider in this work, an overview on camera motion types, and some of the video degradations commonly experienced in user generated video.

3.1 Auxiliary sensors in mobile devices

As we mentioned, one of the main contributions of this work is the exploitation of auxiliary sensor modality for analyzing user generated video content. It is therefore important to introduce the sensors we use:

- Tri-axial accelerometer,
- Compass (tri-axial magnetometer).

A tri-axial accelerometer records acceleration across three mutually perpendicular axes. One very important characteristic of this sensor is that when there is lack of other acceleration it senses static acceleration of $1g$ (approximately 9.8 m/s^2 at sea level) in the direction of Earth's center of mass. This relatively strong static acceleration allows identifying the tilt of the camera with respect to the horizontal plane, i.e., the plane that is perpendicular to the gravitation force. We fix the camera orientation with respect to the three perpendicular accelerometer measurements axes as shown in Fig. 2.

We consider electronic compasses realized from tri-axial magnetometers. These sensors output the instantaneous horizontal orientation towards which they are pointed with respect

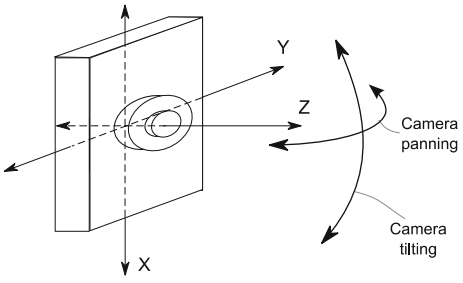

Fig. 2 Illustration of accelerometer data axes alignment with respect to the camera. Also the two camera movements considered in this work are shown (i.e., panning and tilting)

to the magnetic north. That is, the output of the compass is given in degrees from the magnetic north. By using a tri-axial magnetometer, the sensed orientation of the camera is correct even in the presence of tilt (with respect to the horizontal plane).

Accelerometers and compasses provide mutually complementary information about the attitude of the device in which they are embedded. Accelerometers provide the angle of tilt with respect to the horizontal plane whereas compasses provide information about the angle of rotation within the horizontal plane (with respect to magnetic north). In case of a camera embedding these sensors, an accelerometer can provide information about tilt movements, whereas a compass can provide information about panning movements.

We assume that the sensor readings are sampled at a fixed (but possibly different for the individual sensors) sampling rate. Also, we assume that the sampling timestamps for the sensor data are available and they are aligned with the start of video recording. The recorded sensor data can be regarded as a separate data stream. In the evaluation section (Section 6), we show that these assumptions are reasonable and can be readily satisfied without specialized hardware setup.

Other auxiliary sensors which can be found embedded in modern smartphones are gyroscopes and GPS receivers. Gyroscopes estimate the rotational position of the device they are embedded in, by measuring and integrating the angular velocity. GPS receivers provide an estimate of the device location in the form of a geographical coordinate pair. However, in this work we do not consider these two types of sensors for the following reasons. For the use cases addressed in this paper the orientation accuracy provided by the electronic compass (for determining panning movements and the ROI) and the accelerometer (for detecting tilting movements and wrong camera orientation) was considered to be sufficient. Also, our test recording devices did not embed gyroscopes, so we relegated their use to the future work. GPS receivers were not considered as they would provide the location information only when the cameras are outdoors, whereas we target both outdoor and indoor public happenings.

3.2 Camera motion

In cinematography camera motion is regarded as one of the most important types of cinematic techniques. Camera motion can be classified into two main categories: rotational and translational. In this work we consider only rotational camera motion, as it is the most common case for user generated videos. In particular, we focus on *camera panning* and *camera tilting*.

Camera panning is the horizontal rotational movement of a camera, whereas camera tilting consists of rotating the camera in the vertical direction with respect to the horizontal plane (see Fig. 2). Both of these camera movements are commonly performed either for following an object of interest or for changing the focus point (or direction) of the recording.

The detection of camera motion is commonly regarded as *global motion estimation*, as opposed to estimating the motion of objects present in the recorded scene. To obtain information about the camera movements, there are mainly two approaches: content based techniques and motion-sensors based techniques. As already described in the section dedicated to the prior art, the most common techniques are based on video content analysis (such as for example in [1] and [24]). Such methods analyze changes in pixel data between subsequent frames of the video for estimating the displacement of objects with respect to the camera view. Unfortunately, these content based techniques are typically computationally demanding, furthermore their performance could be impaired by the motion of objects present in the recorded scene.

3.3 Video degradations

The degradations that we consider are camera shake and wrong camera orientation (with respect to the horizontal plane). We shall briefly describe them in what follows.

Unintentional camera shake is a notable problem in the most common use case of holding the camera by hands (i.e., not supported by a tripod or a stand). The problem becomes even worse with elapsing of the recording time as the hands get tired. Not limited only to unintentional hand shake, there are often other factors such as shake due to movement (walking or being transported) of the user holding the camera. For example, fast or hectic camera moves are a common characteristic of handheld video recording as the user may point the camera swiftly to a different direction in order to capture unanticipated events or simply due to lost attention. These also fall in the category of shake.

When a video mash-up (or summary) is created by temporally merging video segments captured by multiple cameras, it is important to have all video segments recorded in the same camera orientation, i.e., either all in landscape mode (in which the major axis of the sensor/display is parallel to the horizontal plane) or all in portrait mode (in which the major axis of the sensor/display is perpendicular to the horizontal plane). However, in professional multi-camera productions videos are almost never captured and displayed in portrait mode. Thus, we define "wrong camera orientation with respect to the horizontal plane" as the orientation of the camera when it is in portrait mode. This wrong camera orientation represents yet another problem that is characteristic of non-professional video recording. In fact people are used to record for their own personal video archives without worrying about the camera orientation, because each recorded video is usually watched individually (i.e., one at a time).

4 Multimodal analysis of individual videos

In this section we consider individual user generated videos which have been recorded by a camera-enabled device while auxiliary sensor data was simultaneously captured. For each of such videos we analyze the associated auxiliary sensor data for detecting camera motion and video degradations. Performing such detections is important for example in video summarization or in multi-camera video production, where too fast camera motion or low quality video content are not desirable and should be excluded. Furthermore, we utilize indexed camera movements within other multimodal analysis methods that we present in this paper (generic and specific interesting event detection, mutual distance estimation). The important and novel aspects of this section consist of analyzing auxiliary sensor data associated to individual user generated videos (which have been recorded without any restrictions on the use of the camera-enabled devices) in order to detect camera motion and video degradations.

4.1 Detection of individual camera moves

In this sub-section we develop methods for detecting camera panning and camera tilting in individual user generated videos by means of multimodal data analysis. The novelty introduced by our methods consists of exploiting context sensor data (from the auxiliary sensors considered in Section 3.1) for camera motion detection, in order to overcome the limitations of content-based camera motion estimation (already discussed in Section 2 and in Section 3.2). Our global motion detection methods are computationally efficient, and their performance does not depend on the effect of camera motion on the video-content data as it does instead for the content-based techniques. The individual camera moves that are

detected by our proposed methods represent low-level events and they are the basis for discovering the higher-level information presented in Sections 5.2, 5.3 and 5.4. Before describing the details of our panning and tilting detection algorithms (described in subsections 4.1.2 and 4.1.3), we first discuss the way by which we achieve detection of camera motion that takes into account the camera zoom level.

4.1.1 Zoom-invariant camera motion

It is important to notice that camera movements affect the video content that is being recorded in a different manner depending on the angle of view of the camera. The angle of view is the angular extent of the scene which is captured by the image sensor. A camera move (panning or tilting) of certain angular extent has greater impact on the recorded scene when the angle of view is narrower. The angle of view depends on the size of the image sensor, on the focal length (optical zoom) and on the application of digital zoom. Provided that the image sensor has fixed size, the changes to the angle of view are due to changes in the zoom (optical or digital). Due to non-square aspect ratio of the imaging sensors, the angles of view across the horizontal and across the vertical dimensions differ; thereafter, we refer to them as α_h and α_v, respectively.

The actual change to which the recorded video scene is subject during a panning depends on the angle of view in horizontal direction α_h used during the recording, as illustrated in Fig. 3. For a camera panning of θ_p degrees the camera view changes by a ratio of $r_h = \left(\theta_p/\alpha_h\right)$ where $r_h \geq 1$ indicates a complete change of the viewed scene. For a ratio $r_h < 1$, the panning will result in keeping $100 r_h\%$ of the view after the panning ends. Analogously, for camera tilting, the angle of view in vertical direction affects the actual change of scene ratio, which is given by $r_v = (\theta_t/\alpha_v)$, where θ_t is the span of the camera tilting.

Consequently, the angle of view affects the perceived (from the video content) speed of camera motion, which we analyze for detecting camera panning and tilting. We thereafter propose to measure the scene-change rates R_h and R_v as indicators of the perceived speed of

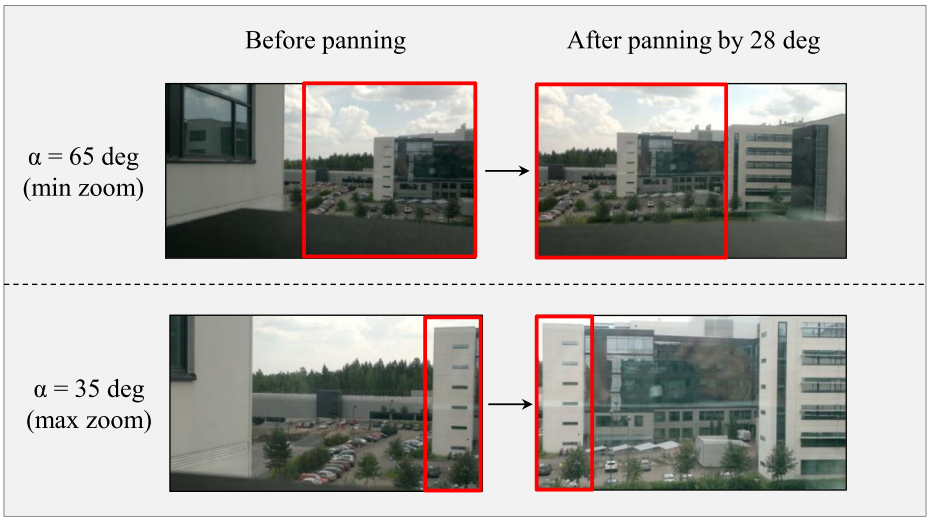

Fig. 3 Illustration of the different visual impact of a panning extent of 28° on the recorded scene, when the zoom level changes

camera motion, which are obtained from the angular speed of the camera (in respectively horizontal and vertical directions) and from the angle of view in the following way:

$$R_h = \frac{r_h}{\Delta t} = \frac{\theta_p}{\alpha_h \cdot \Delta t} = \frac{\omega_h}{\alpha_h} \tag{1}$$

$$R_v = \frac{r_v}{\Delta t} = \frac{\theta_t}{\alpha_v \cdot \Delta t} = \frac{\omega_v}{\alpha_v} \tag{2}$$

where ω_h and ω_v are the angular speeds respectively in horizontal direction and in vertical direction.

4.1.2 Camera panning

In order to detect the camera panning, we analyze the data captured by the electronic compass during the video recording activity, instead of relying on video content analysis. This method for detecting camera panning requires much less computational complexity than methods based on video content analysis. Also, we analyze the real motion of the camera and not the effects the camera motion has on the recorded video content. Therefore, our method is not affected at all by the motion of objects on the scene.

For detecting a panning event we perform the following steps:

Algorithm 1—Camera panning detection

1. Apply low-pass filtering to the raw compass data.
2. Compute the first discrete derivative over time of the filtered compass signal, which represents the angular speed of the camera.
3. Scale the obtained angular speed by the reciprocal of the angle of view, obtaining the rate of scene-change in the horizontal direction R_h.
4. Select the peaks of the obtained rate of scene-change by using a threshold T_P.

The low-pass filtering in the first step is necessary to avoid obtaining peaks which are due to short or shaky camera movements, rather than to intentional camera motion. The threshold has been set as $T_P=3.08\cdot10^{-3}$ based on extensive empirical optimizations and it is such that for the reference value of the angle of view $\alpha_h=65\ deg$ (which corresponds to minimum zoom level in our case), a panning is detected if the camera moves horizontally with angular speed of at least $\omega_h=2\ deg/sec$. An illustration of how camera panning is detected from the compass signal is given in Fig. 4. Each camera panning is represented by two timestamps: start- and stop-panning timestamps.

Furthermore, we classify the detected panning movements with respect to their speed. A panning can be a slow one, performed by the user with the intent of capturing the scene during the camera motion (e.g., for obtaining a *panoramic* panning or for following a moving object), or a faster one, for example performed in order to change the focus point (or direction) of the video recording. As already described in Section 4.1.1, a certain angular speed of a panning can have different visual impacts on the recorded scene when using different angles of view (e.g., different zoom levels). In fact, for the same angular speed the scene recorded during a panning changes more rapidly when the camera is using a greater level of zoom. Therefore, in order to account for the effects of the zoom level, for classifying each detected panning we compare its scene-change rate R_h with a threshold $T_S=9.85\cdot10^{-3}$. If R_h is greater than T_S then we mark the panning movement as fast, otherwise as slow. The

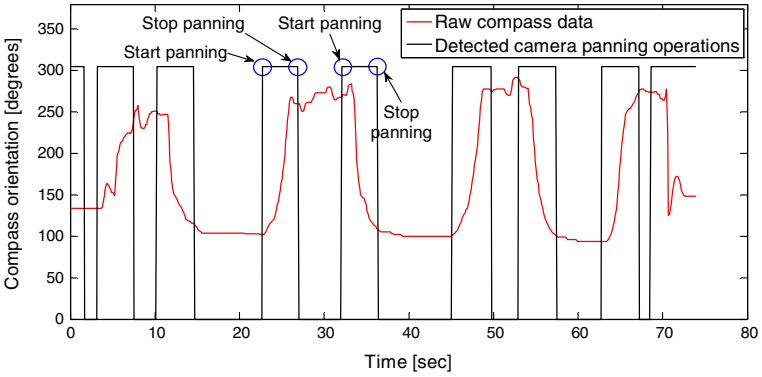

Fig. 4 Detection of camera panning movements

threshold T_S has been set after extensive experimentations and it is such that, for the reference value of the angle of view α_h=*65 deg*, a panning with horizontal angular speed exceeding *6.4 deg/sec* is considered a fast panning. Figure 5 shows a sequence of panning movements and their classification. Furthermore, Fig. 6 shows the detection and classification of panning movements together with the video frames captured before and after each panning.

4.1.3 Camera Tilting

For detecting camera tilt movements, we exploit the ability of the accelerometer to detect both:

- the steady-state camera tilt angle (i.e., when no tilting is being performed by the cameraman),
- the speed and duration of a camera tilting.

We assume that the camera tilting involves marginal translational acceleration—especially when compared to the Earth's gravitational component of $1g$. Actually, this assumption should be satisfied in order to capture video that does not suffer from obtrusive camera shake or motion blur. When the cameraman is performing a tilting movement, the gravitational component changes so that it gets distributed between the X and the Z axes from Fig. 2.

Let us define the angle of tilt as the angle between the Z axis and the horizontal plane. The angle of tilt for a given instantaneous static acceleration vector, $\overrightarrow{A} = [A_X, A_Y, A_Z]$ is

$$\theta = arctan \frac{A_Z}{\sqrt{(A_X)^2 + (A_Y)^2}} \tag{3}$$

We compute the discrete derivative of the obtained instantaneous angles, which indicates the speed at which the tilt angle changes. Let us denote this derivative as D_θ. In order to detect the temporal extent of a camera tilting move, we first compute the vertical scene-change rate R_v as:

$$R_v = \frac{|D_\theta|}{\alpha_v}, \tag{4}$$

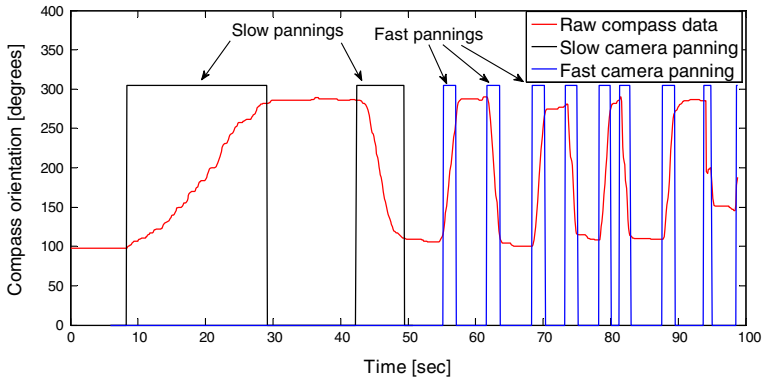

Fig. 5 Classification of camera panning movements based on angular speed of the camera

and then we compare it with a predefined threshold, T_t, and obtain an array, T, whose elements indicate the detected tilting movement:

$$T = 1_{\{R_v > T_t\}} \tag{5}$$

That is, nonzero values of T indicate that a tilting is being performed.

4.2 Detection of video content degradations

This subsection presents methods that exploit context sensor data modality to detect degradations (camera shake and wrong camera orientation) in individual user generated videos. The degradations are detected for the whole duration of a given video. This is accomplished by segmenting the video into small segments of fixed duration. The same segmentation is performed also for the sensor data.

4.2.1 Camera shake

We analyze streams of recorded sensor readings, representing contextual information of a user generated video, to detect camera shake and classify it according to its strength. As a result of the analysis we mark the video segments which have been detected as shaky, in order to be evaluated by multi-camera production systems or multimedia retrieval applications.

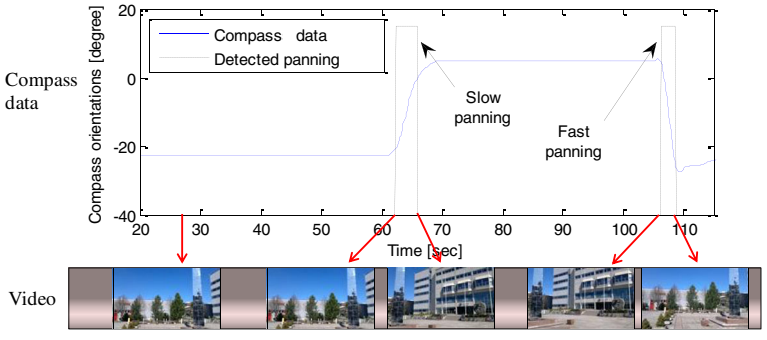

Fig. 6 Illustration of the detected and classified panning movements, and the video frames captured before and after each panning

We propose to analyze changes in the static acceleration across the three axes of a tri-axial accelerometer in order to detect shake. More precisely, we detect changes of the device orientation with respect to the horizontal plane provided that there is relatively small translational acceleration as compared with the static acceleration (an assumption that is satisfied when holding the camera by hands). We assume that the frequency band 10–20 Hz (see [33]) in the accelerometer-data spectrum contains contributions mainly from shake and thereafter we use it for our further processing. Filtering out the lower frequency components is important in order to avoid slower non-shake camera movements, such as panning and tilt, being detected as shake. The overall algorithm is as follows.

Algorithm 2—Camera shake detection

1. Apply a high-pass filter with cut-off frequency of 10 Hz on each of the three components of the accelerometer data (i.e., corresponding to the X, Y, and Z axes).
2. For each video segment, do the following:
   a. Compute the sample variance of each of the three filtered accelerometer components and store as, Ex, Ey, and Ez
   b. Compute the median value of Ex, Ey, and Ez and assign it as the amount of shake in the video segment.

It is worth discussing why step 2b is needed. Even though we assume that the shake affects all three components, a fast tilt or panning can produce a significant contribution to one of these three components—typically stronger than what hand shake contributes. Thus, in order to avoid such moves affecting on the shake estimation, we take the median value of the sample variances as representative of the level of shake. We further classify the shake as *strong* in case the obtained median value is greater than τ_{shake}, which is a predefined threshold. Figure 7 is an illustration of how the detection of shake (and also of wrong camera orientation) is performed.

4.2.2 Wrong camera orientation

The camera orientation (landscape or portrait) can be detected by measuring the static component of the accelerometer (i.e., the gravitational acceleration acting on the device). This fact is being exploited in mobile phones to detect landscape or portrait mode of operation. We make use of it to detect whether video segments are shot with correct or wrong camera orientation, for the reasons that we illustrated in Section 3.3. Without loss of generality, let us assume that the diagram in Fig. 2 shows the correct orientation of the camera with respect to the axes of the accelerometer. This is the case when the static acceleration does not have any contribution along the Y axis. We assume the orientation of the recording device being a relatively slowly-varying event (as otherwise it would fall into the category of shake). Thereafter, we consider the lower frequency portion of accelerometer components (by applying a low-pass filter with the same cut-off frequency as the one used for the shake detection, i.e., 10 Hz). To detect the camera orientation in any of the video segments, we compute the average of the instantaneous orientations within this segment. The instantaneous orientations are computed with a formula analogous to Eq. (3),

$$\theta_Y = arctan \frac{A_Y}{\sqrt{(A_X)^2 + (A_Z)^2}} \tag{6}$$

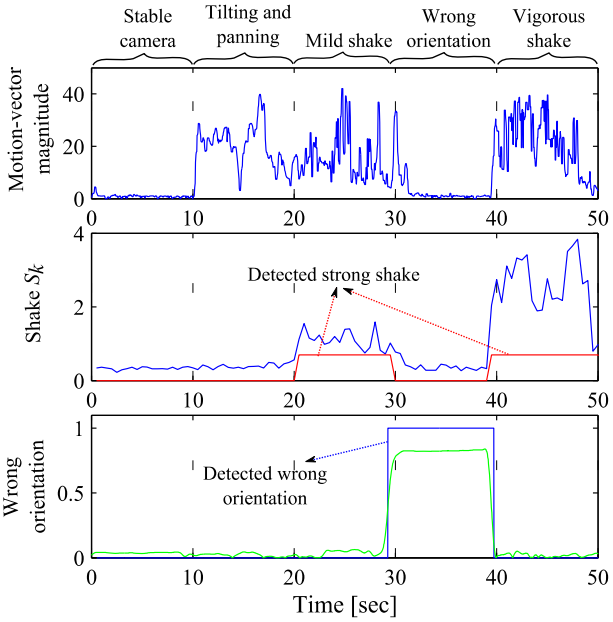

Fig. 7 Detection of video shake and of wrong camera orientation

and then we check for wrong phone orientation by comparing $|\theta_Y| > \theta_{wo}$, where θ_{wo} is a predefined parameter that is the maximum allowed angle of deviation.

5 Multimodal analysis of videos recorded by multiple users

In this section we now consider a recording scenario in which multiple people are simultaneously recording the same happening, for example, a live music show or a sport happening. The most important and novel aspects of the work described in this section are a joint analysis of the multiple recordings and exploitation of auxiliary sensors (providing context data), where both content and context data are captured by mobile phones without any restrictions on their usage. In particular, the outcomes of our analysis techniques are the identification of the region of interest, the detection of generic interesting events, the estimation of the relative distance between users and examples of detecting more specific interesting events. Some of the reasons for performing such analysis lie in the need for understanding high-level information about a public happening which has been recorded by multiple cameras. For example, this information can be subsequently used for indexing the video content, so that it can be easily browsed or retrieved, or for automatically creating video summaries or multi-camera video mash-ups. In particular, determining the region of interest is important for understanding which segments of the recorded videos are the most relevant; extracting interesting/salient events would provide information about the most important video segments; determining the relative distance of cameras is useful for a better awareness of the recording activity. Furthermore, the detection of specific interesting events exploits the knowledge of the identified ROI for inferring higher-level semantic information.

5.1 Identifying a region of interest

In some use cases it is important to know which area of a public happening is considered as the most interesting by the audience. In this work we propose to identify the ROI by exploiting the fact that most public happenings have a single scene of common interest, often this is the performance stage. Most of the people who are recording usually point their camera towards the stage for most of the recording time. Therefore, we consider the time spent by a camera user recording towards a certain direction as an implicit indication of his/her interest on the scene. In addition, we exploit the availability of multiple cameras in order to exclude outlier camera users (i.e., users recording mostly towards an orientation which is interesting only for them). Thus, we identify the ROI as the specific angular range of orientations (with respect to North) towards which the users have recorded video content for most of the time.

As we already discussed in Section 2, prior works have focused on different types of scenarios, mainly on surveillance and environment monitoring. Instead our method attempts to understand the angular region in which people are interested the most in a public happening. The novelty of our ROI estimation method consists of analyzing solely auxiliary sensor data captured by multiple cameras for inferring the ROI. The algorithm works as follows:

Algorithm 3—ROI detection

1. The preferred angular width of the ROI is a parameter to be set manually.
2. For each recording user, compute the time spent recording in each orientation and update the histogram of recorded time units (e.g., seconds) over the orientations.
3. Analyze the obtained histogram and find the angular range (of user-specified width) that maximizes the length of total video content which has been captured towards such a range. This angular range represents the ROI.

The width of the ROI can be set to any reasonable value. In our experiments we have set it to 90° or 180°. Figure 8 shows an example in which seven camera users recorded a live music performance by pointing their cameras towards a limited range of orientations for most of the recording time.

5.2 Generic interesting event detection

During public shows such as live music performances or sports happenings, there are often certain events that attract the attention of some or most of the attending people, and thus also of those who are recording video. We consider such events as *interesting* events. For example, an interesting event in a music show would consist of the guitar player starting playing a solo or moving around on the stage. When such events occur, the most natural thing that a person who is recording would do is to capture the event, by turning the camera towards the subject (e.g., the guitar player) that has triggered the interesting event and eventually by following its movements. As a result, the cameras of these people will be moved in a similar (or correlated) manner; a similar idea is utilized for automatic zooming in [5]. We consider this correlated and simultaneous motion of some cameras as a potentially interesting event for the viewer of the recorded videos, as it was found interesting and worth recording by multiple camera users. In contrast, a camera movement performed solely by one camera user might have been triggered by something which is of interest only for that particular user. As the specific type of event that has triggered the attention of the camera users will not be identified, we denote such events as *generic* interesting events. We detect

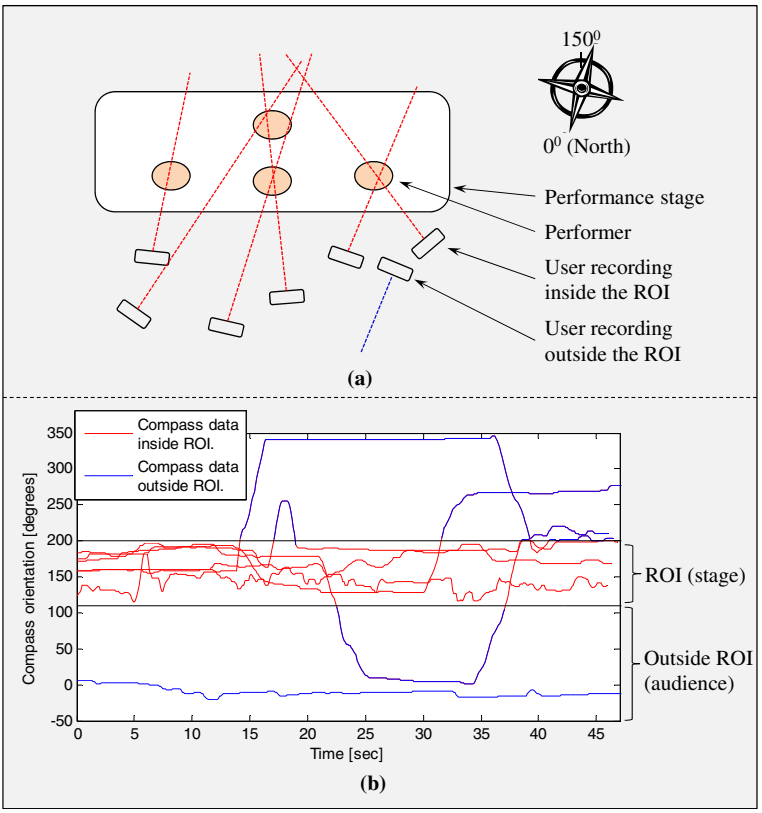

Fig. 8 ROI identification. **a** Scenario where multiple users record a live music performance using their cameras. **b** (*Unwrapped*) compass data captured by the users during the video recording. As can be seen in the plot, six users pointed their cameras towards a similar orientation for most of the recording time. We have specified a preferred ROI width of 90°. The proposed method identified the ROI to be in the range [110, 200] degrees, which is satisfactory, as it corresponds to orientations pointing towards the stage of the recorded happening. During the recording session some of the users performed panning movements either inside or outside the identified ROI. In particular, one user has been recording, all the time, something which was outside the ROI

the camera motion for each individual video recording by means of sensor data analysis as described in Section 4.1. We assume that the recordings of all users are mutually temporally aligned; the alignment itself is a topic on which we do not elaborate further in this paper.

As already discussed in Section 2, many previous works have considered the problem of detecting interesting events either from a single video or from multiple videos. However, most of them analyze exclusively video and/or audio content data, thus being computationally very expensive. The novelty of our approach consists on both the type of data on which the analysis is performed, and the types of interesting events that are detected. In fact, we analyze solely auxiliary sensor data which is a computationally light-weight approach. The analysis of such type of data allows us to exploit the interest implicitly shown by the recording users, which would hardly be possible to determine by performing only content-based analysis.

In the following we describe how we analyze compass data and accelerometer data for detecting generic interesting events from respectively correlated camera panning and correlated camera tilting.

5.2.1 Correlated camera panning

In order to detect generic interesting events from correlated panning movements we perform correlation analysis on the compass data captured by different camera users. In this way, we aim to discover temporal segments within which the compass data from some users is highly correlated.

Thus, it is likely that in those time intervals the users have turned their cameras in a similar way (i.e., similar temporal extent of the camera movement, similar speed, etc.). Figure 9 shows an example in which a number of users is recording a stage and at some point a subset of these users turn their cameras towards an interesting point (the performer 1). For the correlation analysis we have chosen to compute the *Spearman*'s rank correlation coefficient [9], as it allows for detecting similar panning movements even if the cameras are being turned to opposite directions or the speed of the panning varies. After the computation of the correlation has been performed, we apply an empirically set threshold in order to detect the main peaks of correlation. From this thresholding operation we get two timestamps that refer to the beginning and the end of the highly correlated camera panning. They represent the temporal boundaries of the generic interesting events. In order to filter out some irrelevant correlated camera motion (i.e., not related to a substantial change in the orientation), we consider as correlated camera panning each of those temporal segments in which there is co-occurrence of a correlation peak and of camera panning operations performed by the camera users.

The method could be summarized in the following steps:

Algorithm 4—Detection of correlated camera panning

1. Compute the Spearman correlation between the raw compass data streams produced by the cameras.
2. Detect the major peaks of correlation by using a threshold.
3. Detect the camera panning movements for each camera.
4. Detect the co-occurrence of a correlation peak for a certain set of cameras and of panning movements for the same set of cameras.

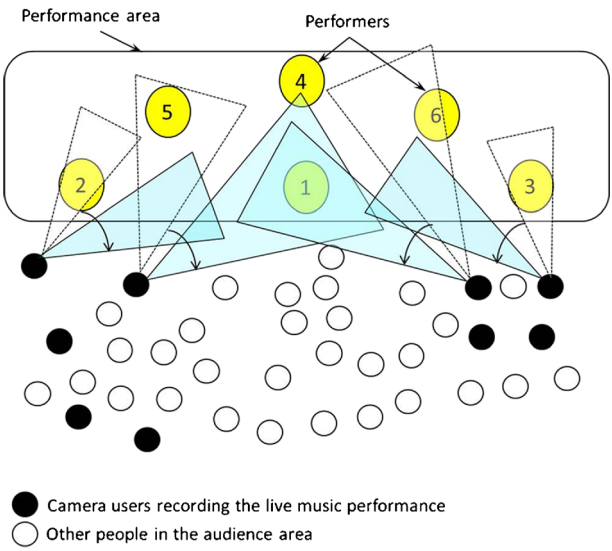

Fig. 9 Correlated camera panning movements (*view from the top*). The orientations of the cameras before and after the correlated panning are represented respectively by the *dotted lines* and the *solid lines*

5.2.2 Correlated camera tilting

In this section we consider correlated camera tilts performed by two or more users by comparing their individual detected tilts as given by Eq. (5). An illustration of this event is given in Fig. 10a, where two users are recording and following an object of interest moving in vertical direction. Let us denote the detected tilt of the k-th user as T_k. Then, we classify two or more camera tilts as correlated if and only if their temporal beginnings and ends are within a predefined threshold, t_{corr}, and the tilt directions are coinciding. The tilt directions are given by the sign of the derivative D_θ. Let us consider the individual tilts of any two camera users (detected as described in Section 4.1), described by their pairs of beginning- and ending- timestamps, t_1^{beg}, t_1^{end} and t_2^{beg}, t_2^{end}. These camera tilts are detected as correlated if and only if $|t_1^{beg} - t_2^{beg}| < t_{corr}$ and $|t_1^{end} - t_2^{end}| < t_{corr}$.

For the correlated tilt detection we use a different approach than for the correlated panning detection as we consider as correlated only tilts performed in the same direction. The reason for this is that camera tilting in opposite directions is highly unlikely to be a

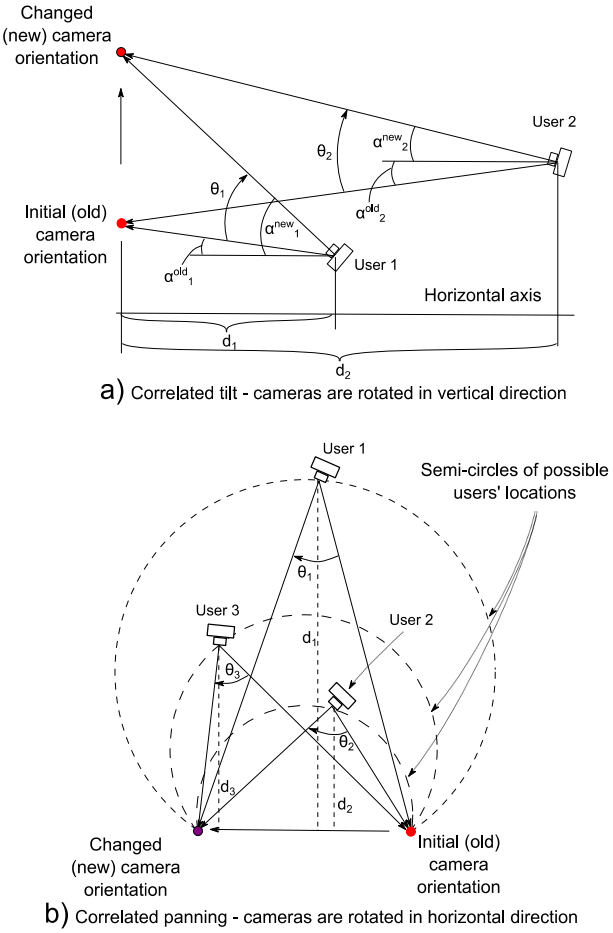

Fig. 10 Illustration of camera users performing correlated camera tilting (**a**) and correlated camera panning (**b**)

result of following the very same object of interest. Such an unlikely scenario is when two camera users are at substantially different altitudes and the object or scene they are recording is positioned between them in altitude. In contrast, camera panning in opposite directions is rather common as the camera users can have various locations with respect to the scene of interest.

5.3 Detecting mutual distance between users

The correlated camera panning and tilting discussed in the previous sections allow us to infer some information about the relative positions of the users that perform these camera movements with respect to the common object (scene) of interest. More precisely, the goal is to detect the ratio of the distance of each user to the common object of interest.

The novelty of our method consists of utilizing only the correlated motion of cameras capturing the same moving object, which is detected solely by auxiliary sensor data analysis. Also, compared to the prior art discussed in Section 2, our method does not make use of any specific stationary landmark for inferring the mutual distance.

We consider separately the methods exploiting correlated panning and tilting since for the correlated tilting we have additional information (the reference of the horizontal plane) and this allows us to detect the exact ratio of distances of the users who perform correlated tilting.

Let us denote the steady camera tilt angles before and after the tilting movement of the k-th user as respectively ${\alpha^{old}}_k$ and ${\alpha^{new}}_k$, where the individual tilts are part of a detected correlated tilting; an illustration is given in Fig. 10a. Also, let us denote the unknown distance from the scene to the k-th user as d_k. After simple calculations, the ratio between the distances d_i and d_j of users i and j is computed as

$$\frac{d_i}{d_j} = \frac{tan\left({\alpha^{old}}_j\right) - tan\left({\alpha^{new}}_j\right)}{tan({\alpha^{old}}_i) - tan({\alpha^{new}}_i)}, \tag{7}$$

where α^{old}, $\alpha^{new} \in (-\pi/2, \pi/2)$ and $\alpha^{old} \neq \alpha^{new}$. In the above equation, the ratio is always non-negative as the correlated tilts are always in the same direction (up or down). In the same manner we can compute the ratio of distances (to the scene) between each pair of users and a map of such distances can be computed. To obtain all distances in absolute values (e.g., in meters), it is required to know the absolute value of only one of the distances (from a user to the moving object), i.e., the absolute distance of at least one user with regards to the commonly viewed scene.

The correlated camera panning can also be used to infer some information about the distance between users. However, it is limited by the fact that the direction in which objects of interest have changed is unknown, as cameras can be also on opposite sides of the object and thus perform panning in opposite directions. Instead for tilting the change is simply in the same vertical direction. Thereafter, for each user we detect a circle on which the user's coordinates can be located; this is illustrated in Fig. 10b. We propose to compute the ratio between the greatest possible distance for each user from the axis connecting the old and the new objects of interest:

$$\frac{\max(d_i)}{\max\left(d_j\right)} = \frac{tan\left(\theta_j/2\right)}{tan(\theta_i/2)}, \tag{8}$$

where θ_j and θ_i are the individual angles of panning. Thereafter, the obtained ratio between the maximum distances can be used as approximation of the actual ratio d_i/d_j.

It is worth noting that, in these methods for estimating the relative distance of users by exploiting correlated motion, we assume that the camera users are not moving (or else their movements are negligible compared to the distances between them). This is a fairly reasonable assumption as performing video recording and simultaneously moving the camera would result in significant camera shake.

5.4 Application of generic interesting events for detecting specific events

The generic interesting events described in Section 5.2 can be used as a building block for discovering other more specific events in which correlated camera motion among users is meaningful, by integrating them with the analysis of other data modalities. In this section we give an example of how the detected generic interesting events can be used for discovering some specific and higher-level events for a particular type of public happenings being recorded—live music performances. The specific event detection algorithm presented here takes advantage of two main ideas. The first one consists of jointly analyzing data streams of same modality (e.g., compass data) captured by multiple devices, in order to detect simple events (each one obtained from one sensor modality). The second idea consists of exploiting the availability of data from multiple modalities, by combining the detected simple events for obtaining higher-level semantics of the media analysis. We consider two examples of such higher-level events that can be defined by means of compound sentences:

S_1: "Music starts and some users make a correlated panning to point to the stage",
S_2: "Music ends and some users make a correlated panning towards the audience".

These events represent something quite common in real-world scenarios of live music shows. In fact, regarding the event S_2, at the end of each song the audience usually starts applauding, and those people who are recording a video in many cases decide to capture such a salient moment by turning their cameras towards the audience. Similarly, it is also common to point the camera back to the stage when the next song starts. We use the following simpler events (that we denote also as primitive events) for detecting the specific events: camera panning correlation, audio class, and audio similarity degree. In the following we will first describe the considered primitive events and then we will provide more details about how we obtain the specific ones.

5.4.1 Primitive event detection

The first primitive event that we consider is the camera panning correlation, already described in Section 5.2.1. In particular, for the specific event S_1 we detect correlated camera panning movements that end towards the identified ROI, whereas for S_2 we consider only those correlated panning movements that end towards directions outside the ROI.

In order to detect the instants when a song starts and ends (i.e., for performing song segmentation) we classify the audio into "Music" and "No music" using the audio classifier described in [17], which we have specifically trained for live music shows. The end of class "No music" and the start of class "Music" represent the start of a song, whereas the end of class "Music" and the start of class "No music" represent the end of a song. As the audio classification is applied on live music audio recorded by mobile phones, the presence of ambient noise and the quality of the microphones reduce the performances of the audio classifier. As a result, we obtain a relatively high rate of missed detections for the start or end of songs. To overcome this problem, we detect start and end of songs also with an additional method that exploits the

following idea. While a song is being played, even if the cameras are spread all over the audience area, they will capture similar audio content, which is output by the loudspeakers. On the contrary, after the song has ended, the cameras will not capture the same sound anymore and the audio similarity among them will likely be low. We take advantage of this for performing song segmentation by determining whether the audio content recorded by the users is similar or not. We interpret the transition in the audio similarity degree for most of the cameras from low to high and from high to low as respectively the start and the end of a song. We compute the similarity degree for temporally overlapping audio segments extracted from the video recordings of different cameras. For computing the audio similarity, we use a recent work which is described in [17]. Then we apply a threshold on the output of the analysis in order to discriminate between low and high similarity degree.

5.4.2 Combining primitive events for detection of specific events

For detecting the two specific events S_1 and S_2, we consider the co-occurrence (within a predetermined temporal window) of a panning correlation peak and a state change for at least one of the two song segmentation methods. S_1 is detected when a panning correlation peak (for some cameras) occurs (and the panning movements end towards the ROI) and at the same time there is a transient from no-song to song (detected either by the audio classifier or by the audio similarity algorithm, or both). Then we can assert that the song has started and very likely some of the users have pointed their cameras towards something of interest within the ROI, probably the performance stage. S_2 is detected when there is a co-occurrence of a panning correlation peak (and the panning movements end towards orientations outside the ROI) and a transient from song to no-song (detected either by the audio classifier or by the audio similarity algorithm, or both). Then we assert that the song has ended and very likely some of the users have pointed their cameras to something they found interesting outside the ROI, for example the audience which is applauding. In case the two song segmentation methods provide conflicting results (i.e., for the same temporal window one method detects the start of a song and the other detects the end), then we consider only the detection provided by the audio classifier, as it has proved to be more reliable than the audio similarity analysis for the song segmentation task.

6 Evaluation

In this section we evaluate the performance of the proposed methods: detection of individual camera moves, detection of video degradations, ROI estimation, detection of generic interesting events, estimation of mutual distance between cameras, and examples of detecting more specific events. As our methods mainly analyze streams of sensor measurements captured simultaneously with the video recording, there are no publicly available datasets that already contain such sensor data. Therefore, we use special test data which is collected and partitioned into two datasets as described in Section 6.1. For capturing the test data we used publicly available high-end smart phones and simple dedicated software that collects the sensor data synchronously with video recording. The default sampling rate for each sensor was used, i.e., 40 samples per second for the accelerometer and 10 samples per second for the compass, where each sample is further labeled with its timestamp. The angular resolution of the compass is 1°. The time alignment between a recorded video and the recorded sensor data was straightforward to obtain, as the video start- and stop-recording times were obtained from the creation time of the media file and were then matched to the

timestamps of the sensor measurements. The software application that we used stored the auxiliary sensor measurements (and associated timestamps) as data streams associated with the recorded video. In particular the sensor data streams were stored in text/plain format, as there are no standardized formats for storing this type of data together with the associated video content.

In this section we use the following measures for evaluating the performances of the detection methods:

- *Precision* (*P*), computed as the ratio *true positives/(true positives+false positives)*, where *true positives* are the correct detections and *false positives* are the non-correct detections.
- *Recall* (*R*), computed as the ratio *true positives/(true positives+false negatives)*, where *false negatives* are the missed detections.
- *Balanced F-measure* (*F*), computed as the ratio $(2 \cdot P \cdot R)/(P + R)$. It is the harmonic mean of the precision and the recall.

At the end of this section (in subsection 6.8) we provide a discussion on the experimental results for all tested methods.

6.1 Test datasets

We used two datasets in our experiments: dataset D_1 which was obtained in a real world scenario, and dataset D_2 which was obtained in a simulated scenario.

Dataset D_1 contains 47 video recordings from two different live music shows (featuring 10 bands), spanning an overall time duration of about 62 min. The data was collected by nine users that were attending the shows and were sparsely located in the audience (so that there is no visual interference between them during the recording). The age of the users varied between 25 and 40. They had different video recording skills: some of them (about the half) were used to record using a mobile phone, whereas the others only rarely used their phones or other devices for recording videos. Furthermore, most of the users did not know in advance the songs being played at the show (i.e., they were not fans of the bands). The participants were not given any specific instructions on the way of recording. On the contrary, they were only asked to record the happening as they would normally do for their own personal use. The only constraint was to use camera-enabled phones that we provided (instead of their personal phone), thus a short explanation about how to access the camera application was given to the users beforehand. Dataset D_1 has been used for evaluating the following methods: detection of individual camera moves, detection of video degradation, estimation of the ROI, detection of generic and specific interesting events.

In order to obtain a dataset with significant number of camera movements, for dataset D_2 we simulated a recording experiment in a large auditorium where two camera users were asked to record what they were considering interesting for a duration of 25 min. During that time, moving objects were displayed in different parts of the large auditorium display, serving as potential reason to perform a camera move. Even though the users were used to utilize mobile phones for their everyday life, they were not used to record videos with them (i.e., only occasionally). Dataset D_2 has been used for validating the following analysis methods: detection of individual camera moves, detection of video degradations, detection of generic and specific interesting events, and estimation of the mutual distance between cameras.

Datasets D_1 and D_2 are not publicly available for legal reasons, mainly due to copyrights of the bands playing in the live music shows (for dataset D_1) and in the simulated public happening (for dataset D_2).

6.2 Detection of individual camera moves

In Section 4.1 we described a method which detects camera movements by solely analyzing auxiliary sensor data captured during the video recording. In particular, camera panning and camera tilting are detected by analyzing data captured by the electronic compass and by the accelerometer, respectively.

For evaluating the performance of the camera motion detection methods the following approach has been adopted. Firstly we visually inspected the recorded video content contained in datasets D_1 and D_2 in order to manually annotate the occurrence of camera panning and tilting movements. Furthermore, each panning event was manually labeled as either slow or fast depending on whether the video content captured during each panning was respectively pleasant or unpleasant to watch. Afterwards, we applied the proposed detection methods on the datasets. The obtained results are summarized in Table 1. We also applied our method for panning speed classification on the detected panning movements. Table 2 summarizes the performance of the proposed panning classification method on datasets D_1 and D_2. In particular, in the table we report the classification accuracy only for those panning movements that were correctly detected (i.e., only for the true positives).

6.3 Detection of video degradations

Two methods for detecting video degradations by solely analyzing auxiliary sensor data were described in Section 4.2. In particular, we proposed to detect camera shake and classify its strength, and to detect the temporal segments during which a camera is being held in portrait mode (considered as wrong orientation for the reasons explained in Section 3.3).

We have tested the algorithms using datasets D_1 and D_2 which include segments of normal camera usage (standing still or performing slow or faster camera movements), segments with camera shake, and segments where the camera was held in a wrong orientation (i.e., portrait mode). The approach for evaluating our proposed methods is as follows. By visual inspection we manually annotated 522 video segments with shake. Out of these, 495 were correctly identified by the proposed shake detection technique. In 43 segments instead our method has erroneously detected shake. Thus, we achieved a precision of 92%, a recall of 94.8% and a balanced F-measure of 93.4%. We compare the proposed shake detection method with global motion estimation based on content-analysis block matching. We specifically selected a fully static scene with significant amount of details in order to set

Table 1 Test results for detecting camera movements (panning and tilting) performed by individual users. *gt* stands for *ground truth* (manually annotated events), *tp* stands for *true positives*, *fp* for *false positives*, *p* for *precision*, *r* for *recall*, and *f* for *balanced f-measure*

Dataset	GT	TP	FP	P	R	F
Camera panning						
D_1	129	112	12	0.90	0.87	0.89
D_2	104	104	2	0.98	1.0	0.99
Total	233	216	14	0.94	0.93	0.94
Camera tilting						
D_1	73	69	7	0.91	0.95	0.93
D_2	207	204	9	0.96	0.99	0.97
Total	280	273	16	0.94	0.97	0.96

Table 2 Test results for the classification of panning movements based on speed. The accuracy is computed as the ratio between the number of correctly classified panning movements and the total number of panning movements which were correctly detected (true positives)

Dataset	Confusion matrix			Accuracy (%)
D_1		Automatically detected as fast	Automatically detected as slow	96.4
	Manually annotated as fast (ground truth)	25	1	
	Manually annotated as fast (ground truth)	3	83	
D_2		Automatically detected as fast	Automatically detected as slow	97.1
	Manually annotated as fast (ground truth)	32	0	
	Manually annotated as slow (ground truth)	3	69	
Total		Automatically detected as fast	Automatically detected as slow	96.8
	Manually annotated as fast (ground truth)	57	1	
	Manually annotated as slow (ground truth)	6	152	

an ideal condition for the content-based global motion estimation. From Fig. 7, we observe that in regions that do contain shake, the proposed sensor-based shake detection produces results that are also in line with the results of the global motion estimation. However, the global motion estimation also erroneously triggers when otherwise non-shake camera movements are performed, whereas the proposed method does not suffer from that. At the thirtieth second of the video presented in Fig. 7 it is possible to see how false positives are caused also by camera tilting and by rotating the phone. In the bottom plot of the same figure, we show also the detection of a camera being turned to portrait mode; the detected wrong camera orientation is accurate and follows well the change in static tilt angle associated with the device orientation (shown in green in the same plot). The wrong phone-orientation was detected correctly in all 51 cases when there was such; misclassification of vigorous shake as wrong phone orientation occurred for 2 video segments.

6.4 Identification of region of interest

In Section 5.1 we described a method for identifying the (angular) region of interest in a public happening by inferring the attention of the people recording the happening itself. In fact, we determine the horizontal angular range towards which most of the people are pointing their cameras for most of the recording time. For obtaining the horizontal orientation of each camera we analyzed the compass data.

For evaluating the proposed ROI identification technique we adopted the following approach. Firstly, for each recorded public happening, we manually estimated the angular range where the stage of the musical performance was located with respect to the audience area. For this, we identified the concert area (both stage and audience areas) on a satellite map which provided also the reference direction of North. Afterwards, we estimated the

range of orientations with respect to North (i.e., by considering North as at orientation of 0°, in the same way as electronic compasses do) of lines departing from different points within the audience area and ending into different points within the stage area. In particular, we selected the range in order to consider an ROI width of 90°. This manually identified range represented the ground truth in our evaluation. After this manual estimation, we applied our algorithm on the test data (i.e., the compass data captured by all cameras recording the public happening) and then compared the automatically identified ROI with the one that we had manually estimated.

Regarding dataset D_1, for the two live music shows our method identified the ROIs as the ranges [279°, 9°] (see Fig. 11) and [40°, 130°]. By comparison with the ground truth we verified that the identified ROIs are correct, i.e., they correspond to the actual ranges of orientations towards which the two performance stages were positioned with respect to the respective audience areas. We did not evaluate the ROI identification method on dataset D_2 because the use of only two recording cameras would not be statistically significant for such a task which is based on the exploitation of where most of the users point their cameras.

6.4.1 Combined use of camera panning and region of interest

Figure 12 shows an example of using the detected and classified panning movements, and the identified ROI for automatically generating a multi-camera video mash-up. The data for this example is taken from one of the two live music performances contained in dataset D_1. For the sake of simplicity, in this figure we consider only two cameras. For each camera, the plot of the compass data captured during the video recording is shown. Some relevant frames extracted from the recorded videos are shown below the associated compass data. These frames have been extracted at time instants indicated by the red arrows (mostly before and after each panning which is relevant for this example). The bottom-most part of the figure represents the multi-camera video mash-up obtained by stitching together segments extracted from the two source videos. These segments are delimited by the camera-switches instants. In the video mash-up a switch to a certain camera is triggered by the occurrence of a panning (which ends within the ROI) performed by that camera. The switching times are indicated by the vertical dashed lines that originate either from video

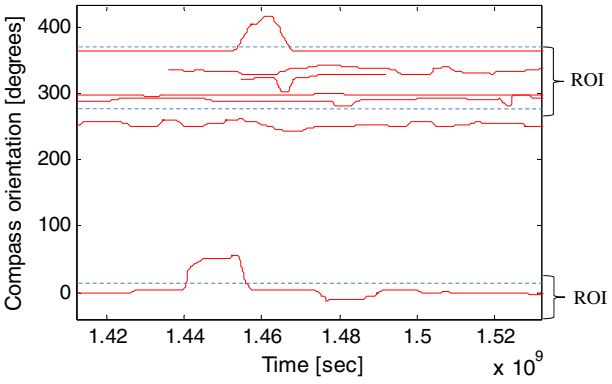

Fig. 11 Excerpt of (*unwrapped*) compass data captured by multiple cameras in one of the live music shows in dataset D1. For this happening the ROI has been identified in the range [279°, 9°] when a preferred ROI width of 90° was specified

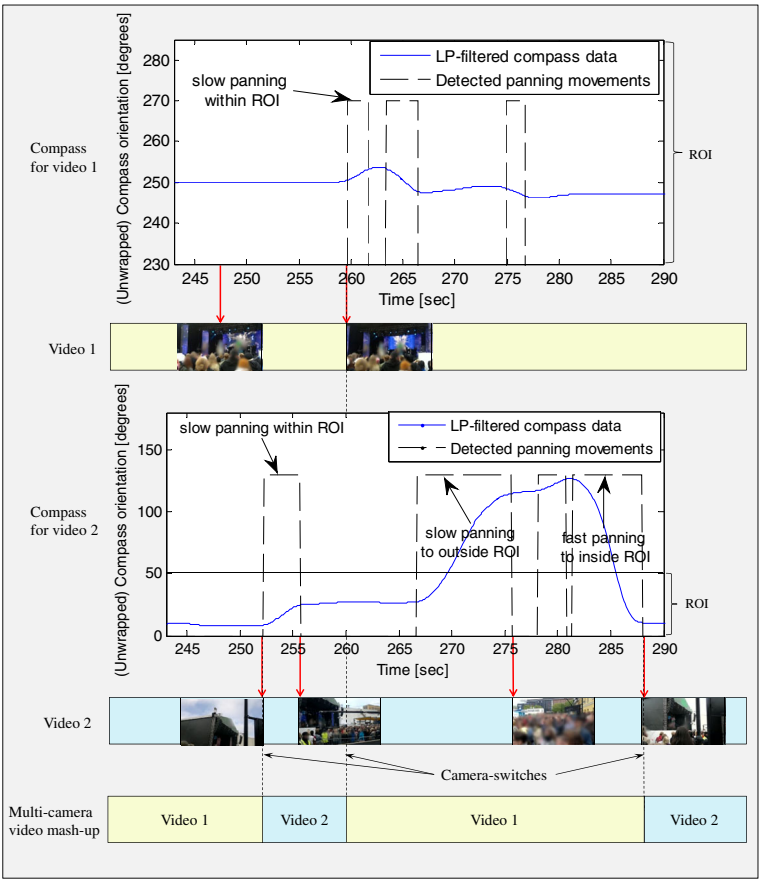

Fig. 12 Combined use of panning movements, their classification and Region of Interest for automatically generating a multi-camera video mash-up

1 or video 2 (either at the beginning or at the end of each panning movement), depending on which camera has triggered the switch. Panning detection, panning classification and ROI identification has been performed by using our proposed methods. For this example a preferred ROI width of 180° has been chosen; the identified ROI is in the range [230°, 50°]. The user of camera 1 performed several slow panning movements, recording always within the ROI. The user of camera 2 performed three slow and one fast panning movements. Among the slow movements, the first one is within the ROI, the second one starts inside the ROI and ends outside the ROI, and the third is within the area outside the ROI. The fast panning was performed from outside the ROI to inside.

Due to prior state, the excerpt of the mash-up starts with video 1. At 252 s user of camera 2 performs a slow panning. As the movement ends within the ROI, a camera-switch is performed in the video mash-up from video 1 to video 2. Also, as it is a slow panning, the switch happens at the beginning of the camera motion. At almost 260 s camera 1 performs a slow panning within the ROI, therefore a camera-switch from video 2 to video 1 is triggered. At about 267 s camera 2 starts a slow panning, but since the movement ends towards outside the ROI, no camera-switch is triggered in the mash-up. Another short panning happens in video 2 at 278 s, but again outside the ROI. At about 282 s camera 2 performs a fast panning

that starts outside the ROI and ends inside it, so a switch from video 1 to video 2 is triggered. The actual switching time corresponds to the end of the panning, as the movement is fast (and also because the scene recorded outside the ROI should not be shown in the video mash-up).

6.5 Detection of generic interesting events

A novel approach to detecting generic interesting events was presented in Section 5.2. We proposed to analyze context sensor data to detect temporal segments during which a salient event has occurred. In particular, we detect the time interval during which some people have moved their camera in a correlated way, as this is considered an indicator of common interest. At this stage of the analysis we do not aim at understanding the specific reason of the correlated movements. We detect generic interesting events from correlated camera panning and from correlated camera tilting.

In this section we provide empirical results and their evaluation about these methods. The evaluation has been performed in the following way. By visually inspecting the videos recorded at each public happening, we manually annotated events which are commonly considered as interesting in the context of live music performances, such as the start of singing, the start of a guitar solo, or something funny/unusual that happened either on the performance stage or within the audience area. For each manually extracted event we noted down the (absolute) timestamps of its beginning and end. Afterwards, we ran the detection algorithms on the test data and then compared the obtained results to the ground truth by checking whether the time intervals of the automatically and of the manually detected events were coinciding or within a predefined temporal window. Furthermore, for considering a detected event as a true positive we performed a further validation step. We considered those cameras that were detected as performing correlated movements and in particular the video-content segment that they recorded during the detected time interval. By visual analysis we verified that the cameras actually recorded the manually extracted salient event.

6.5.1 Generic interesting events from correlated camera panning

We applied the detection of generic interesting events from correlated camera panning (described in Section 5.2.1) on recordings of real live music performances (dataset D_1) and on recordings in simulated environment (dataset D_2). For dataset D_1, we manually annotated 50 interesting events. Following is a list of some types of detected interesting events for that experiment:

- Audience putting hands up or clapping.
- Person being lifted by the audience.
- Start of singing.
- Start of guitar solo.
- Guitar player playing hi-hat (a cymbal).
- Performers moving across the stage.
- Music starting after a break.

Figure 13 shows the detection of some interesting events from dataset D_1. The experimental results are summarized in Table 3. Based on these results we can assert that the proposed method is effective in recognizing such interesting events when recording real live music performances.

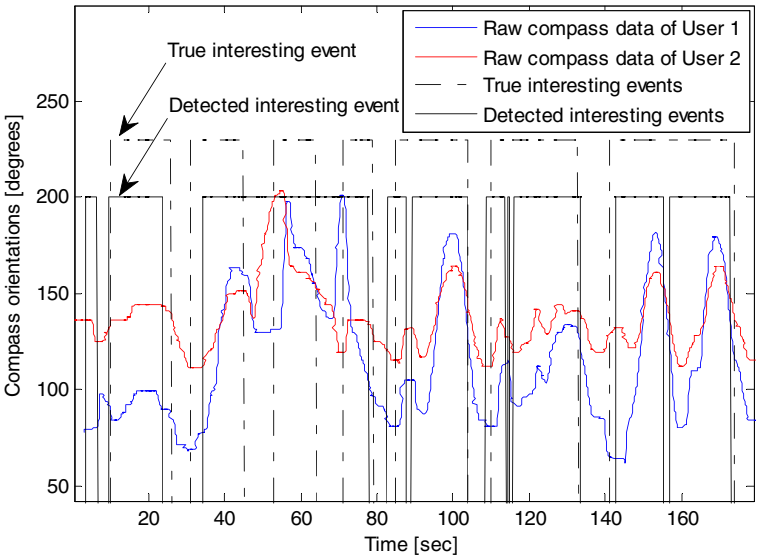

Fig. 13 Correlated camera panning movements

6.5.2 Generic interesting events from correlated camera tilting

The detection of interesting events from correlated camera tilting movements is obtained by applying the method proposed in Section 5.2.2. The objective performance evaluation of this method is given in Table 3. Dataset D_1 contains much fewer correlated tilting movements as compared with correlated panning; one explanation is that changes in horizontal direction are more probable for capturing an object of common interest. Plots of the detected camera tilt angles and the detected correlated tilts are illustrated in Fig. 14. In this plot, it can be seen that false positive detection of individual camera tilts (seen in the plot for both camera users 1 and 2) are not resulting in a false positive for the correlated camera tilt—as there is no matching in both users.

Table 3 Test results for detecting generic interesting events. *gt* stands for *ground truth* (manually annotated events), *tp* stands for *true positives*, *fp* for *false positives, p* for *precision*, *r* for *recall*, and *f* for *balanced f-measure*

Dataset	GT	TP	FP	P	R	F
Generic interesting events from correlated panning						
D_1	50	43	17	0.72	0.86	0.78
D_2	58	48	4	0.92	0.83	0.87
Total	108	91	21	0.81	0.84	0.83
Generic interesting events from correlated tilting						
D_1	5	5	1	0.83	1.0	0.91
D_2	112	105	9	0.92	0.94	0.93
Total	117	110	10	0.92	0.94	0.93

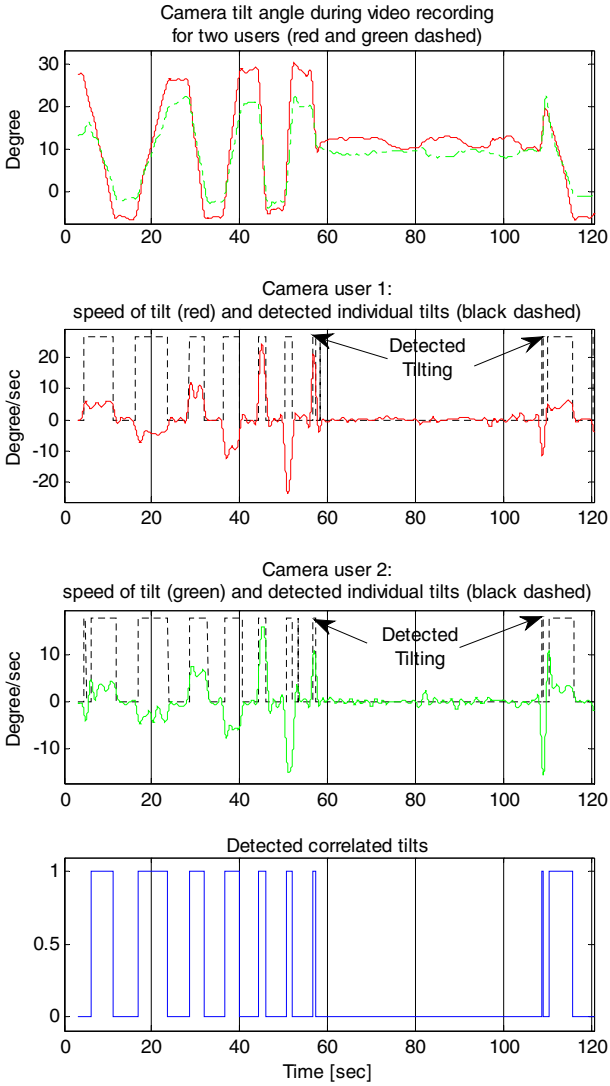

Fig. 14 Example of detecting interesting events from correlated camera tilting

6.6 Detection of relative users' distances to a common point of interest

In Section 5.3 we proposed methods for estimating the mutual distance between cameras, which analyze only auxiliary sensor data. In particular, we consider those temporal segments in which correlated camera movements between different cameras occur.

In order to evaluate our proposed methods specified by Eqs. (7) and (8), we applied them on dataset D_2 for detecting mutual distances from correlated camera movements. We note that for dataset D_1 we could not obtain ground truth user locations since the users had been changing their locations throughout the performance—thereafter, we could not use this dataset to validate the proposed methods. For dataset D_2, we had two different cases (known to us); in the first case, the distance (to the stage) of one of the user was twice as much as the

distance of the other and in the second case the two users had exactly the same distance to the stage. The proposed methods based on correlated tilting and panning provided rather faithful estimation of the ratio of the users' distances to the stage when applied to 170 correlated camera movements from dataset D_2. The mean of all estimated distance ratios resulted within 2% of the actual ones (meaning lack of bias) and the sample standard deviation was 10.7% of the actual distance ratios. The value of the standard deviation shows that the estimation errors were relatively small. In fact, three outliers were contributing the most for its value, having errors that were greater than 50% of the actual distance ratio. These outliers were due to the fact that the initial (i.e., just before the camera movement had been performed) points of interest for the users were not the same.

6.7 Detection of specific events

In this section we evaluate the performances of the specific event detector presented in Section 5.4. The method that we proposed consists of combining events which are detected from individual modalities in order to detect more specific and semantically richer events. We provided two examples, S_1 and S_2, of how this can be realized in practice. In these examples the following three primitive events need to be combined: correlated camera panning, audio class, audio similarity degree. In particular, S_1 is detected when there is co-occurrence of correlated panning, audio class changing from "No music" to "Music", and audio similarity degree changing from "Low" to "High". Instead, for detecting S_2 we need to observe the co-occurrence of correlated panning, audio class changing from "Music" to "No Music" and audio similarity changing from "High" to "Low".

For evaluating the performance of this example methods for detecting specific events we used dataset D_1. Dataset D_2 could not be used as this dataset does not include recordings of real live music shows, while events S_1 and S_2 are valid only within the domain of such types of public happenings. For collecting the ground truth we visually inspected the videos and observed the occurrence of both S_1 and S_2 once for each song performed by the bands recorded in dataset D_1, respectively at the beginning and at the end of the songs. For each observed instance of S_1 and S_2 we noted down its timestamps. We then applied the detection algorithm on the test data and for the actual validation of the results we followed a similar procedure as for the detection of generic interesting events. We verified that the detected specific events were temporally coinciding with the manually extracted events. Also, we verified that the cameras detected as performing correlated movements in the manual annotation and in the proposed algorithm were actually the same.

The test results and the performance measures are given in Table 4. Figure 15 shows an example (extracted from the second live music show contained in dataset D_1) in which the specific event S_2 is detected from two camera users recording one of the test songs.

Table 4 Experimental results for the detection of specific events. *tp* stands for *true positives*, *fp* for *false positives, p* for *precision*, *r* for *recall*, and *f* for *balanced f-measure*

Specific event	Ground truth event	TP	FP	P	R	F
S_1	19	15	4	0.79	0.79	0.79
S_2	19	16	3	0.84	0.84	0.84
Total	38	31	7	0.82	0.82	0.82

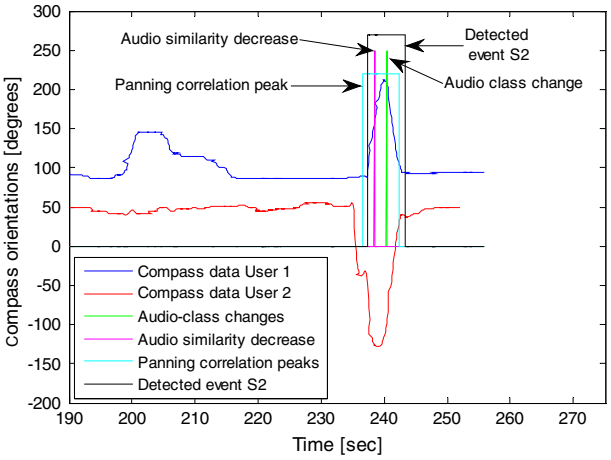

Fig. 15 Detection of specific event S2: audio-class changes from "Music" to "No music", audio similarity degree changes from "High" to "Low", users perform correlated panning

6.8 Discussion on the achieved results

Through extensive experimentations on both real world and simulated scenarios we have shown the effectiveness of the multimodal analysis methods that we proposed in this paper. Regarding the results for the detection of individual camera movements we achieved very high accuracy (expressed as the balanced F-measure) in both the considered datasets (on average, 0.94 for panning and 0.96 for tilting). In addition, the operations involved for detecting such movements are rather simple (mainly low-pass filtering and first-order derivative), especially because they are applied on data which is sampled at very low rates (in total 50 samples per second). Prior works on detecting such movements have mainly focused on analyzing the video content, like in [1] where the authors consider the particular case of user generated videos. Content-based approaches are computationally expensive, as for example 1 s of HD video at 25 frames per second contains about 46 million pixels, whereas in our experiments 1 s of accelerometer and compass data contains only 50 sensor measurements. By looking at the results given in Table 1, we can observe that detection accuracies are higher for dataset D_2. This was expected, as this dataset has been collected in a controlled environment where we have simulated a public happening, thus external disturbances (like big crowds of people, etc.) were absent. In addition, we have tested the classification of panning movements into two classes, slow and fast, obtaining an overall accuracy of 96.8%. The proposed classification method is based on a predefined speed threshold which was set in such a way that video content recorded during a fast panning would not be pleasant to watch. Such classification could be used for example in the automatic generation of a summary or of a mash-up, so that video content recorded during too fast camera movements can be excluded from being used.

We have also evaluated our proposed methods for detecting video segments containing degradations by solely analyzing motion sensor data. Regarding the detection of shake, we achieved good recall performance (94.8%). The obtained false negatives (i.e., missed detections) were due to segments with rather mild shake. We also obtained several false positives (i.e., mistakenly detected shake segments), which were also contributed by camera tilting and by rotating the phone. Also the shake detection, in the same way as the detection

of fast camera movements, can be used for the purpose of excluding some video segments which are not pleasant to watch, for example in video summaries or video mash-ups. The detection of wrong camera orientation has proved to work effectively too, as it detected correctly all ground truth wrong orientation segments, and the number of false positives was very limited.

Regarding the estimation of the angular region of interest we have shown how our method is effective in exploiting the orientation of the cameras recording the public happening. The method is straightforward and it is light-weight in terms of computational complexity, as only compass data (captured by all cameras) is being analyzed. For example, as 1 s of recorded video has 10 compass samples associated to it, the total amount of compass data captured during a show of 1 h and recorded by 10 different cameras would be only 360,000 samples, which is negligible when compared to the amount of video or audio data captured during the same time span.

We evaluated also our proposed method for detecting generic interesting events from correlated panning and correlated tilting. The methods performed well on datasets D_1 and D_2, achieving respectively 0.83 and 0.93 of balanced F-measure. Therefore, the performance for dataset D_1 is lower than for dataset D_2 and, in particular, for D_1 we obtained an increased number of false positives, caused by inadvertent camera movements. By applying these methods we were able to automatically detect events in a way which would have been hard for content-based techniques. One example is the event of a person in the audience which was lifted by other people (this is often observable especially in rock and metal concerts). The prior work discussed in [3] has also addressed the problem of exploiting auxiliary sensor data for detecting interesting events, but the setup used by the authors consists of cameras attached to the clothes of users, which is not the typical use case when attending public happenings such as live music performances. Instead, in our tests we did not put any restrictions on the use of the camera-enabled mobile phones. Figure 16 shows the detection of few generic interesting events for one of the two live music shows in dataset D_1. Also, as an example of usage of the detected events, the automatic generation of a video summary is illustrated, wherein each detected interesting event is included.

Regarding the estimation of relative distance between cameras we showed that our method obtains small errors; this is a good achievement especially when considering that we do not make use of any indoor/outdoor positioning system. Instead we only analyze streams of compass and accelerometer measurements, thus leading to a light-weight approach compared to other methods that analyze audio and/or video data (such as in [2] and [11]). Furthermore, our method is totally independent of the place (i.e., indoors or outdoors) where the distance estimation needs to be performed, as it only exploits correlated motion between multiple cameras. Thus, our approach is very suitable to the use case of analyzing videos recorded in public happenings (e.g., live music performances), which are often held indoors where positioning systems such as the GPS could not be utilized for estimating the distance between cameras.

Finally we evaluated the detection of two example specific events for the domain of live music shows. We obtained an overall balanced F-measure of 0.82 which is a good result, given the semantically high level of the information being inferred with these two specific events (for example S_2 consists of the cameras being turned in a correlated way for recording the audience at the end of a song). Also these specific events would require high computational cost for being detected by means of video content analysis, as for example face detection and other visual analysis methods would be required.

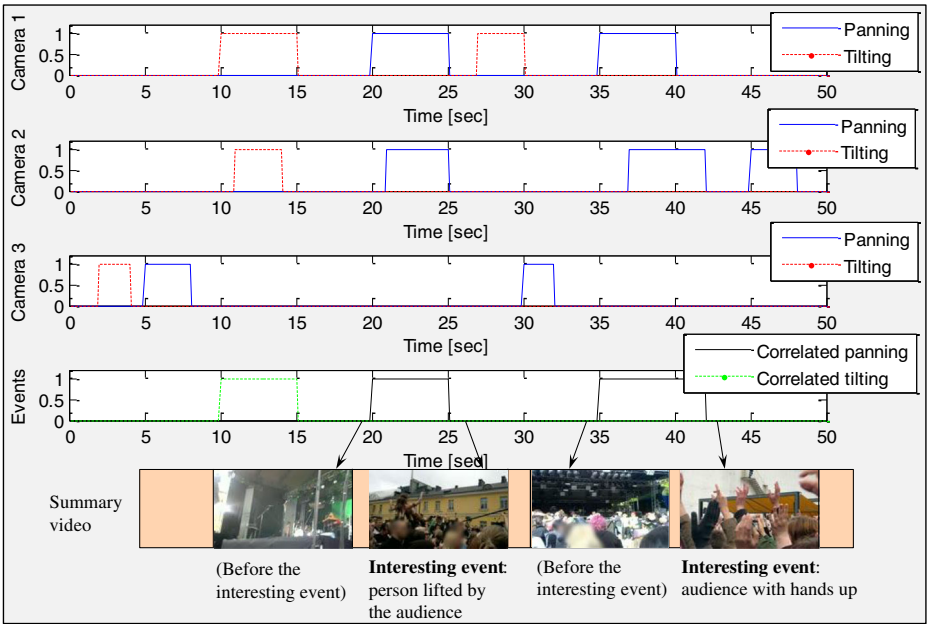

Fig. 16 Illustration of generic interesting events (from correlated panning and tilting) and an example of their usage for video summarization. This test data has been extracted from dataset D_1 (for visualization purposes, only three cameras are shown)

7 Conclusions

In this paper we presented a set of multimodal analysis methods for detecting interesting events and obtaining high-level contextual information in user generated videos.

First, we consider individual videos. By exploiting auxiliary sensor measurements captured during the video recording we are able to detect camera movements (tilting and panning). Previous work (such as [1]) has addressed the problem of detecting camera motion in user generated video by analyzing the video content data. Unfortunately, content-based approaches are prone to errors due to the fact that they measure only the effects of motion on the video data, not the camera movement itself; thus other types of motion, such as motion of objects in the scene, may negatively affect the performance of content-based techniques. Instead our method is independent of motion of objects as it analyzes movements detected by the auxiliary sensors (compass and accelerometer). Furthermore, the use of such sensors avoids the computational burden of content analysis. For individual videos we detect also video content degradations (camera shake and wrong camera orientation) by analyzing accelerometer data which is stored as data streams associated to the video content. This way, a system which automatically generates a video summary or a multi-camera video mash-up can decide to exclude video segments which were detected as being of low quality.

In addition we consider a scenario in which multiple camera users record videos of a common scene (e.g., a live music show). We exploit the availability of multimodal data captured by multiple cameras for inferring information about the recording activity and the recorded scene. By analyzing the compass data captured by all cameras we estimate the angular region of interest of the public happening. This technique is computationally efficient and through experimentations we have shown that it gives satisfactory results.

We analyze auxiliary sensor data captured by multiple cameras also for discovering generic interesting events, by detecting correlated camera movements among multiple users. We evaluated these event detection methods against our datasets and we were able to perform detection of events that would have been hard to achieve by analyzing only content data from a single camera. When considering the detection of high-level events from user generated videos, the proposed methods are among the very first to exploit the auxiliary sensor modality. To our knowledge, only [3] has previously considered this modality for the purpose of video summarization. However, the setup used in that work consists of mobile phone cameras attached to the clothes of some users—whereas in our work we rely solely on user generated content without any restrictions on the camera use. The generic interesting events discovered by our methods are used also for estimating the relative distance of those cameras performing correlated movements. We have shown that this technique performs accurately on test data. Finally, by combining sensor data analysis with audio content analysis, we provide an example in which more specific events (i.e., with higher-level semantics) are discovered.

Our experiments on real and simulated scenarios show that the proposed multimodal analysis methods perform well in extracting the considered events and contextual information. We would like to mention that other sensors can be employed in multimodal event detection. For example, a tri-axial gyroscope can provide a finer resolution of the rotational movements of the camera as compared with the compass and the accelerometer. However, a gyroscope by itself cannot provide a reference with respect to actual space coordinates and thus cannot replace any of these two sensors.

As the conclusions of this work we want to remark that due to the availability of information describing the actual scene and the recording activity from different modalities (e.g., motion sensors describe the device attitude, audio content describes the audio-scene, video content describes the visual-scene, etc.), multimodal analysis can provide results with higher-level semantics (as the ones presented in Section 5.4). The *multi-user* availability and the *multimodal data* availability are jointly exploited in this work and constitute one of its main contributions.

References

1. Abdollahian G, Taskiran CM, Pizlo Z, Delp EJ (2010) Camera motion-based analysis of user generated video. In: IEEE Transactions on Multimedia, vol. 12, no. 1, pp. 28–41. IEEE
2. Anjum N, Taj M, Cavallaro A (2007) Relative position estimation of non-overlapping cameras. In: IEEE International Conference on Acoustics, Speech and Signal Processing, Honolulu, Hawaii, U.S.A., pp II-281–II-284. IEEE
3. Bao X, Choudhury RR (2010) MoVi: mobile phone based video highlights via collaborative sensing. In: 8th International Conference on Mobile Systems, Applications and Services, San Francisco, U.S.A., pp 357–370. ACM
4. Broilo M, Basso A, Zavesky E, De Natale FGB (2010) Unsupervised event segmentation of news content with multimodal cues. In: 3rd International Workshop on Automated Information Extraction in Media Production, Firenze, Italy, pp 39–44. ACM
5. Carlier A, Charvillat V, Ooi WT, Grigoras R, Morin G (2010) Crowdsourced automatic zoom and scroll for video retargeting. In: ACM International Conference on Multimedia, Firenze, Italy, pp 201–210. ACM
6. Cheng AJ, Lin FE, Kuo YH, Hsu WH (2010) GPS, compass, or camera?: Investigating effective mobile sensors for automatic search-based image annotation. In: ACM International Conference on Multimedia, Firenze, Italy, pp 815–818. ACM
7. Dong Y, Hon WK, Yau DKY (2007) On area of interest coverage in surveillance mobile sensor networks. In: IEEE International Workshop on Quality of Service, Evanston, Illinois, U.S.A., pp 87–90. IEEE

8. Drahansky M, Orsag F, Hanacek P (2009) Accelerometer based digital video stabilization for security surveillance systems. In: Security Technology—Communications in Computer and Information Science, vol. 58, pp 225–233. Springer
9. Fieller EC, Hartley HO, Pearson ES (1957) Tests for rank correlation coefficients. Biometrika, vol. 44, no. 3/4., pp 470–481
10. Foote J, Cooper M, Girgensohn A (2002) Creating music videos using automatic media analysis. In: ACM International Conference on Multimedia, Juan les Pins, France, pp 553–560. ACM
11. Huster A, Rock SM (2001) Relative position estimation for manipulation tasks by fusing vision and inertial measurements. In: MTS/IEEE Oceans International Conference and Exhibition, Honolulu, Hawaii, U.S.A., vol. 2, pp 1025–1031. IEEE
12. Jain R, Sinha P (2010) Content without context is meaningless. In: ACM International Conference on Multimedia, Firenze, Italy, pp 1259–1268. ACM
13. Järvinen S, Peltola J, Plomp J, Ojutkangas O, Heino I, Lahti J, Heinilä J (2009) Deploying mobile multimedia services for everyday experience sharing. In: IEEE International Conference on Multimedia and Expo, Cancun, Mexico, pp 1760–1763. IEEE
14. Kaiser R, Torres P, Höffernig M (2010) The interaction ontology: low-level cue processing in real-time group conversations. In: ACM International Workshop on Events in Multimedia, Firenze, Italy, pp 29–34. ACM
15. Kennedy L, Naaman M (2009) Less talk, more rock: automated organization of community-contributed collections of concert videos. In: 18th International Conference on World Wide Web, Madrid, Spain, pp 311–320. ACM
16. Kim ET, Han JK, Kim HM (1997) A Kalman-filtering method for 3D camera motion estimation from image sequences. In: IEEE International Conference on Image Processing, Washington, DC, U.S.A., vol. 3, pp 630–633. IEEE
17. Lahti T (2008) On low complexity techniques for automatic speech recognition and automatic audio content analysis. Doctoral Thesis. Tampere University of Technology
18. Liang CK, Tsai CH, He MC (2011) On area coverage problems in directional sensor networks. In: IEEE International Conference on Information Networking, Kuala Lumpur, Malaysia, pp 182–187. IEEE
19. Money AG, Agius H (2008) Video summarisation: a conceptual framework and survey of the state of the art. In: Elsevier Journal of Visual Communication and Image Representation, vol. 19, issue 2, pp 121–143. Elsevier
20. MPEG-7, ISO/IEC 15938, Multimedia Content Description Interface. http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=34228
21. Ogino R, Suzuki T, Nishi K (2008) Camera-shake detection and evaluation of image stabilizers. In: International Conference on Instrumentation, Control and Information Technology (SICE Annual Conference), Tokyo, Japan, pp 528–531. SICE
22. Ong E, Lin W, Lu Z, Yang X, Yao S, Pan F, Jiarrg L, Moscheni F (2003) A no-reference quality metric for measuring image blur. In: International Symposium on Signal Processing & Applications, Paris, France, vol. 1, pp 469–472. IEEE
23. Park HS, Lim S, Min JK, Cho SB (2008) Optimal view selection and event retrieval in multi-camera office environment. In: IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems, Seoul, Korea, pp 106–110. IEEE
24. Peyrard N, Bouthemy P (2003) Motion-based selection of relevant video segments for video summarisation. In: IEEE International Conference on Multimedia and Expo, Baltimore, U.S.A., vol. 2, pp 409–412. IEEE
25. Poppe C, De Bruyne S, Van de Walle R (2010) Generic architecture for event detection in broadcast sports video. In: 3rd International Workshop on Automated Information Extraction in Media Production, Firenze, Italy, pp 51–56. ACM
26. Rasiwasia N, Pereira JC, Coviello E, Doyle G (2010) A new approach to cross-modal multimedia retrieval. In: ACM International Conference on Multimedia, Firenze, Italy, pp 251–260. ACM
27. Shin SH, Yeo JY, Ji SH, Jeong GM (2011) An analysis of vibration sensors for smartphone applications using camera. In: International Conference on ICT Convergence, Seoul, Korea, pp 772–773. IEEE
28. Shrestha P, de With PHN, Weda H, Barbieri M, Aarts E (2010) Automatic mashup generation from multiple-camera concert recordings. In: ACM International Conference on Multimedia, Firenze, Italy, pp 541–550. ACM
29. Shrestha P, Weda H, Barbieri M, de With PHN (2010) Video quality analysis for concert video mashup generation. In: Lecture Notes in Computer Science, vol. 6475, pp 1–12. Springer
30. Snoek CGM, Worring M (2005) Multimodal video indexing: a review of the state-of-the-art. Springer Multimed Tool Appl J 25:5–35, Springer

31. Velipasalar S, Brown LM, Hampapur A (2006) Specifying, interpreting and detecting high-level, spatio-temporal composite events in single and multi-camera systems. In: IEEE Conference on Computer Vision and Pattern Recognition Workshop, Washington, DC, U.S.A., p 110. IEEE
32. Wu S, Ma YF, Zhang HJ (2005) Video quality classification based home video segmentation. In: IEEE International Conference on Multimedia and Expo, Amsterdam, The Netherlands, pp 1–4. IEEE
33. Yu HC, Liu TS (2007) Design of a slim optical image stabilization actuator for mobile phone cameras. In: Physica Status Solidi, vol. 4, no. 12, pp 4647–4650. Wiley
34. Zakhor A, Lari F (1993) Edge-based 3-D camera motion estimation with application to video coding. In: IEEE Transactions on Image Processing, vol. 2, no. 4, pp 481–498. IEEE

Francesco Cricri received his M.S. degree in telecommunications engineering from University of Trento, Italy, in 2008. He is currently a PhD student at the Department of Signal Processing of Tampere University of Technology, Finland. His current research interests are on multimedia analysis, signal processing, semantic web, and machine learning.

Kostadin Dabov received the M.Sc. and the Ph.D. degrees in signal processing from Tampere University of Technology in 2006 and 2010. Currently, he is a research fellow at the Department of Signal Processing at Tampere University of Technology. His research interests include image and video processing, sensor signal processing, cross-modal multimedia analysis, and efficient design and realization of signal processing algorithms. He is a recipient of the 2010 Young Author Best Paper Award of the IEEE Signal Processing Society.

Igor Curcio received his Ph.D. degree in Signal Processing from Tampere University of Technology (Finland). After 11 years working as freelance consultant and trainer, in 1998 he joined Nokia Corporation, where he hold several research and management positions in the areas of real-time mobile media. He is now Research Leader at Nokia Research Center. He has been active in several standardization organizations (such as 3GPP, DLNA, ARIB) where he covered sub-working group and task forces Chair positions, and (co-) authored over 200 standardization contributions in the areas of adaptive media, VoIP, QoE and transport protocols. Dr. Curcio holds more than 150 international patents and several patents are currently pending. He is an ACM member since 1990 and IEEE senior member since 2004, and has also published over 50 papers in the areas of Mobile Media Applications and Services. Dr. Curcio served in the organizing committees and TPCs of IEEE CCNC, IEEE PerCom, IEEE WoWMoM, IEEE ICCCN, IEEE PIMRC, IEEE ISCC, ACM MUM. His current interest areas include mobile video applications and services, streaming media, Mobile TV and broadcasting, P2P multimedia, social media, multimodal sensing, context-based applications and future media solutions.

Sujeet Mate is a senior researcher at Nokia Research Center, Tampere, Finland. He received his B.E. degree in electrical engineering from REC Surat, India and M.S. degree in electrical engineering from the University of Texas at Dallas. His research interests include real-time and context enhanced multimedia applications. His research interests include crowd-sourced media, multimodal sensing, mobile multimedia, social television and video conferencing.

Moncef Gabbouj received his BS degree in electrical engineering in 1985 from Oklahoma State University, Stillwater, and his MS and PhD degrees in electrical engineering from Purdue University, West Lafayette, Indiana, in 1986 and 1989, respectively.

Dr. Gabbouj is an Academy Professor with the Academy of Finland since January 2011. He is currently on sabbatical leave at Purdue University, West Lafayette, Indiana, USA. He was Professor at the Department of Signal Processing, Tampere University of Technology, Tampere, Finland. He was Head of the Department during 2002–2007. Dr. Gabbouj was a visiting professor at the American University of Sharjah, UAE, in 2007–2008 and Senior Research Fellow of the Academy of Finland in 1997–1998 and 2007–2008. His research interests include multimedia content-based analysis, indexing and retrieval, nonlinear signal and image processing and analysis, voice conversion, and video processing and coding.

Dr. Gabbouj is *Honorary Guest Professor* of Jilin University, China (2005–2010). Dr. Gabbouj is a Fellow of the IEEE. He served as Distinguished Lecturer for the IEEE Circuits and Systems Society in 2004–2005, and Past-Chairman of the IEEE-EURASIP NSIP (Nonlinear Signal and Image Processing) Board. He was chairman of the Algorithm Group of the EC COST 211quat. He served as associate editor of the *IEEE Transactions on Image Processing*, and was guest editor of Multimedia Tools and Applications, the European journal *Applied Signal Processing*. He is the past chairman of the IEEE Finland Section, the IEEE Circuits and Systems Society, Technical Committee on Digital Signal Processing, and the IEEE SP/CAS Finland Chapter. He was also Chairman of CBMI 2005, WIAMIS 2001 and the TPC Chair of ISCCSP 2006 and 2004, CBMI 2003, EUSIPCO 2000, NORSIG 1996 and the DSP track chair of the 1996 IEEE ISCAS. He is also member of EURASIP Advisory Board and past member of AdCom. He also served as Publication Chair and Publicity Chair of IEEE ICIP 2005 and IEEE ICASSP 2006, respectively.

Dr. Gabbouj was the Director of the International University Programs in Information Technology (1991–2007) and vice member of the Council of the Department of Information Technology at Tampere University of Technology. He is also the Vice-Director of the Academy of Finland Center of Excellence SPAG, Secretary of the International Advisory Board of Tampere International Center of Signal Processing, TICSP, and member of the Board of the Digital Media Institute. He served as Tutoring Professor for Nokia Mobile Phones Leading Science Program (2005–2007 and 1998–2001). He is a member of IEEE SP and CAS societies.

Dr. Gabbouj was the supervisor of the Best Student Paper Award from IEEE International Symposium on Multimedia, ISM 2011. He was recipient of the 2005 Nokia Foundation Recognition Award and co-recipient of the Myril B. Reed Best Paper Award from the 32nd Midwest Symposium on Circuits and Systems and co-recipient of the NORSIG Best Paper Award from the 1994 Nordic Signal Processing Symposium. He is co-author of 450 publications.

Dr. Gabbouj has been involved in several past and current EU Research and education projects and programs, including ESPRIT, HCM, IST, COST, Tempus and Erasmus. He also served as Evaluator of IST proposals, and Auditor of a number of ACTS and IST projects on multimedia security, augmented and virtual reality, image and video signal processing.