Hindawi Publishing Corporation
Advances in Multimedia
Volume 2012, Article ID 292064, 17 pages
doi:10.1155/2012/292064

Research Article

Multimodal Semantics Extraction from User-Generated Videos

Francesco Cricri,[1] Kostadin Dabov,[1] Mikko J. Roininen,[1] Sujeet Mate,[2] Igor D. D. Curcio,[2] and Moncef Gabbouj[1]

[1] *Department of Signal Processing, Tampere University of Technology, P.O. Box 553, 33101 Tampere, Finland*
[2] *Nokia Research Center, P.O. Box 1000, 33721 Tampere, Finland*

Correspondence should be addressed to Francesco Cricri, francesco.cricri@tut.fi

Received 1 November 2011; Revised 17 February 2012; Accepted 12 March 2012

Academic Editor: Wei-Ta Chu

User-generated video content has grown tremendously fast to the point of outpacing professional content creation. In this work we develop methods that analyze contextual information of multiple user-generated videos in order to obtain semantic information about public happenings (e.g., sport and live music events) being recorded in these videos. One of the key contributions of this work is a joint utilization of different data modalities, including such captured by auxiliary sensors during the video recording performed by each user. In particular, we analyze GPS data, magnetometer data, accelerometer data, video- and audio-content data. We use these data modalities to infer information about the event being recorded, in terms of layout (e.g., stadium), genre, indoor versus outdoor scene, and the main area of interest of the event. Furthermore we propose a method that automatically identifies the optimal set of cameras to be used in a multicamera video production. Finally, we detect the camera users which fall within the field of view of other cameras recording at the same public happening. We show that the proposed multimodal analysis methods perform well on various recordings obtained in real sport events and live music performances.

1. Introduction

The widespread use of camera-enabled mobile devices has allowed people to record anything that they find interesting in their daily life. In particular, one of the most popular means for recording videos is represented by mobile phones which, thanks to their easy portability, are available at any time of the day. Interesting things that people consider worth capturing are very diverse; examples may include funny moments with friends or with the family, music shows, celebrations such as weddings. In particular, there are some situations in which a multitude of people happen to be recording the same scene at the same time. These situations are usually public happenings such as sport events or live music performances. In this paper, we target such kind of scenarios, in which videos of the same event are recorded by multiple people for their own personal archives using their handheld devices (we use the terms *happening* and *event* interchangeably).

As also stated in [1, 2], user-generated videos are then seldom watched either by the people who have shot them or by others. One of the main reasons is the lack of effective tools for automatically organizing the video archives in such a way that it would be easy for a user to retrieve a particular video. For example, it would be beneficial to automatically *classify* videos according to genre (i.e., sport, music, travels, etc.), scene (i.e., indoors versus outdoors, cityscape versus landscape), type of venue where the event is held (e.g., stadium-like venues).

Applications targeting video browsing or automatic creation of video summaries would benefit from the availability of salient information about the videos, such as salient events (e.g., a goal that was scored during a football match), and salient regions (e.g., the goal area).

Video recordings captured by multiple cameras at the same event can be utilized for automatically generating a *multicamera video mash-up* (i.e., a temporal sequence of video segments recorded by different cameras and stitched together one after the other) or a *multicamera summary*. These kinds of applications would benefit from the availability of several types of information, such as which cameras provide the best views in terms of some specified quality

measures, or where other cameras are positioned with respect to one specific recording camera.

In this work, we perform multimodal analysis of videos recorded by multiple users at a public happening in order to extract information for indexing the recorded content. The obtained indexing can then be utilized for automatically organizing video archives into classes or for automatically generating multicamera video mash-ups and summaries.

We propose methods for classifying the type of recorded event according to the following criteria:

(i) *indoor versus outdoor event*, by utilizing the GPS lock status information from all the recording devices;

(ii) *event genre* (sport versus live music): we propose novel multimodal features (i.e., features derived from auxiliary sensor data) which are used, in combination with content-based features, to classify the event genre by means of machine learning techniques;

(iii) *event layout* (stadium versus nonstadium): for this we analyze the way by which cameras are spatially distributed and oriented (i.e., the structure of the camera network).

Furthermore we developed methods which identify the following aspects in a multicamera recording scenario:

(i) the *area of interest* within the event area, by exploiting the locations of the recording devices and the way by which they are pointed by their users;

(ii) the *optimal cameras* to be used for automatically producing a multicamera video mash-up;

(iii) the *cameras in the field of view* of other recording cameras.

The novelties which are common to all the methods proposed in this paper are mainly two.

(1) We analyze *contextual data* solely or in combination with video and audio content data. Such contextual data is captured by *auxiliary sensors* (which are embedded within the recording devices) during the video recording activity. In particular we consider data captured by accelerometers, electronic compasses, and GPS receivers.

(2) We exploit the availability of *multiple devices recording the same event* for increasing the robustness of the analysis (thanks to higher redundancy) and for inferring semantic information which would otherwise be hard to extract by analyzing only a single video.

The paper is organized as follows: Section 1.1 introduces the auxiliary sensors used in this work, Section 2 presents the prior works for each of the proposed algorithms, Section 3 describes our proposed methods, Section 4 presents the experimental evaluation, Section 5 is a discussion on the achieved results, and Section 6 concludes the paper.

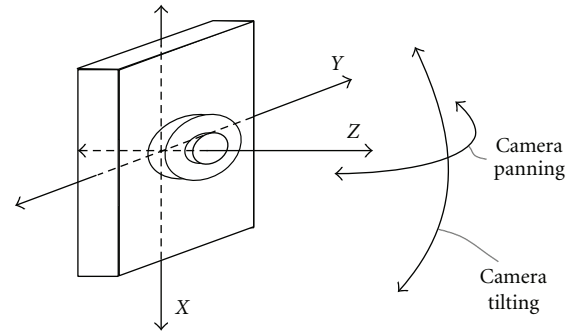

Figure 1: Camera movements and data axes used by the accelerometer and compass; the X-axis is perpendicular to the horizontal plane.

1.1. Auxiliary Sensors in Mobile Devices. Since one of the main contributions of this work is the exploitation of auxiliary sensor modality for analyzing user-generated video content, it is important to introduce the sensors we use:

(i) GPS receiver,

(ii) accelerometer,

(iii) compass (triaxial magnetometer).

Nowadays GPS receivers are present in many electronic devices. For example they are embedded in most modern smartphones, as they are used by those mobile applications which require the location information, such as maps, weather widgets, and image geo-tagging functionality.

A triaxial 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 1 g (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, that is, 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 Figure 1.

We consider electronic compasses realized from triaxial magnetometers. These sensors output the instantaneous horizontal orientation towards which they are pointed with respect to the magnetic north. That is, the output of the compass is given in degrees from the magnetic north. By using a triaxial magnetometer, the sensed orientation of the camera is correct even in the presence of tilt (with respect to the horizontal plane). In case of a camera embedding these sensors, 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 Section 4, we show that these assumptions are reasonable and can be readily satisfied without specialized hardware setup.

2. Prior Art

In this section we report on prior works addressing problems which are similar to those considered by our proposed methods. For each of these works we describe the approach, the type of data which is analyzed, the main differences with respect to our methods, and what are the advantages and drawbacks of our approaches with respect to the prior art. In particular, we focus on works addressing the classification of videos (based on indoor/outdoor scene, on the genre, and on the layout of the recorded event), the identification of the area of interest, the selection of optimal cameras, and the detection of cameras which fall within the field of view of other cameras.

2.1. Classification of Videos according to Indoor versus Outdoor Scene. Many authors have previously worked on analyzing video content for the purpose of classification. A survey on this topic is given in [3]. Regarding the classification of video into indoor/outdoor scene, Serrano et al. [4] proposed an efficient method in which a two-stage classification approach using Support Vector Machines is applied on low-level color and texture features. The authors report accuracy results which are comparable to other more computationally expensive methods. Recently, Lipowezky and Vol developed an indoor/outdoor detector which is suitable for mobile phone cameras [5]. The proposed method works on the Bayer domain image and uses photometrical and colorimetrical features which are normally computed in mobile phones for white balance gains evaluation. The classification step is based on gentle boosting. In [6] the authors propose to use the following features for indoor/outdoor scene classification of images: histograms of Ohta color space, multiresolution simultaneous autoregressive model parameters, and coefficients of a shift-invariant DCT. The method results in 90.3% of correct classification. Payne and Singh [7] propose a method for indoor/outdoor image classification by analyzing the straightness of edge contours in an image. They assume that indoor images in general have larger proportion of straight edges compared to outdoor images. The method recognizes outdoor scenes with strong natural elements and indoor images with structural edges clearly visible, but has problems with urban outdoor scenes and cluttered indoor images. Thus this work would have some limitations when applied on videos of outdoor public happenings, which are often held in urban areas.

All these works address the problem of scene classification by analyzing content data, which is a computationally expensive approach, even though attempts to decrease the complexity have been done, as it is shown for example in [5]. In our method we do not analyze video or audio content at all and instead we only rely on the GPS receiver data provided by multiple recording devices which are present at the event.

2.2. Classification of Videos according to the Genre. Various approaches to the classification of video based on genre have been proposed in the past—using mostly video content analysis. In [8] the authors propose to use domain-knowledge independent features (in particular Scale Invariant Feature Transform) and a bag-of-visual-words-(BoVW-) based model with an innovative codebook generation. For the final classification a k-nearest neighbor classifier is adopted. The method was tested on videos of 23 different sports; therefore it aimed mainly at categorizing subgenres of the sport video genre. The work presented in [9] deals with the use of a hierarchical ontology of video genres. Visual spatio-temporal features are extracted from videos and they are classified using hierarchical Support Vector Machines. In particular the authors propose to construct two optimal SVM binary trees, local and global, in order to find the best tree structure of the genre ontology. The extracted temporal features are average shot length, cut percentage, average color difference, and camera motion, whereas the spatial features are face frames ratio, average brightness and average color entropy. We want to point out that some of these features, namely, average shot length and cut percentage, could not be applied for analyzing user-generated video which is usually unstructured and unedited. It is worth noting that the authors mention that music videos are characterized by larger frame difference (in terms of color histograms), which is a feature that we take into account in our event genre classifier. The proposed method is tested on TV recordings. In [10] the authors discriminate among five video genres—cartoon, commercial, music, news, and sport—by exploiting a combined model of extracted features which are categorized into editing (shot boundary changes), color (color histogram, average brightness, and average saturation), texture (statistics extracted from the gray level cooccurrence matrix, contrast, homogeneity energy, entropy, and correlation), and motion (brightness change, peacefulness of the video, dynamic feature in RGB space). The classifier used is the modified Directed Acyclic Graph Support Vector Machine model. In [11] multimodal features are extracted from TV programs and classified by a parallel neural network into seven genres (commercials, news, weather forecasts, cartoons, music, talk show, and football). The extracted features are color, texture, motion, average shot length, shot cluster duration and saturation, shot length distribution, shot temporal activity, face position distribution, covering percentage of faces, face number distribution, audio segmentation analysis, background audio analysis, and average speech rate. The authors report a classification accuracy rate of 96%. In [2] a genre classification for home videos is proposed, which is of particular interest to us as we target nonprofessionally produced video content too. The authors extract low-level features from MPEG compressed domain claiming that these features are robust to low production quality, which is a common case for home videos. The authors target only those video genres which are specific to home videos, that is, travel, sports, family and pets, event, and entertainment. The extracted features are camera motion (by analyzing motion vectors), subject motion, audio class, audio volume, luminance, color, and flashlight. The authors report that by using ensemble learning they achieve F-measure values of about 0.7 to 0.8.

As we have already mentioned, the discussed genre classification methods analyze video or audio content by

extracting usually complex features. Also, for most of these works the authors consider professionally recorded video content, which is very different from user-generated content. Apart from content data, in our genre classification method we analyze also data captured by the electronic compass and the accelerometer for extracting camera motion features. In this way we avoid performing content-based motion estimation which is computationally expensive and its performance is limited by the presence of moving objects in the recorded scene.

2.3. Classification of the Event Layout. As for our knowledge, no previous works have addressed the specific issue of classifying the type of venue in which a public event is taking place. However, there are some works that address similar issues and they are all based on content analysis, like the previously discussed prior art on indoor/outdoor scene classification and on genre classification. A recent paper ([12]) presents an interesting approach for location recognition. The authors use Speeded-Up Robust Feature (SURF) descriptors to detect objects in images. Location recognition is done by matching the detected objects and their spatial relations between query and database images. However, their approach is aimed for matching images of same exact location rather than classifying between different location types. Schroth et al. [13] give a detailed description of a close-real-time mobile server-based visual location recognition system. They use Maximally Stable Extremal Regions (MSER) as feature detector, Speeded-Up Robust Feature (SURF), and Compressed Histograms of Gradients (CHoG) as key-point descriptors, and the Bag-of-Features (BoF) model for forming the overall descriptor. Also in this case, their visual content-based approach considers matching exact locations rather than location types. Apart from these approaches for location recognition based on image matching, other authors have dealt with the problem of determining the structure of a network of visual sensors, which is what we perform for achieving the type of venue classification. In [14] the authors proposed to measure the statistical dependence between observations in different cameras. In one of the tests they performed, two cameras are positioned at two nonoverlapping portions of a road. In another test five cameras belonging to a real traffic network were used. As also the authors state, the obtained results are approximate but promising. The authors also propose a method for learning the absolute locations of cameras by exploiting the information given by a GPS-enabled device which moves through the recorded area.

As opposed to these content-based approaches, we propose to infer the type of venue by only analyzing the location and the orientation of the cameras. We achieve this by utilizing the data provided by the GPS receiver, the accelerometer, and the compass. These sensors directly provide, respectively, the location, vertical orientation, and horizontal orientation, which would be hard to estimate by means of content analysis only.

2.4. Area of Interest. Analysis of the area or region of interest for multiple temporally aligned video recordings of a common scene has been addressed in several previous works, such as in [15, 16]. In [15] Thummanuntawat et al. use a codebook of local visual features extracted from group of pictures (GOP) formed from the frames of the different views. The resulting features are used with spatial and appearance models as well as motion and depth estimation to track the regions of interest in the scene. Hayet et al. use local image features extracted from video content of multiple cameras to track players in a soccer match [16]. They use modular system architecture and distributed computing to compensate for the high computational cost of local feature extraction and multitarget tracking. Carlier et al. [17] propose a crowd-sourced approach for determining regions of interest (ROI) in a video. They collect usage patterns from a zoomable web video player and consider the regions on which the viewers zoom in as the ROIs. They use Gaussian Mixture Models (GMM) to model the ROIs from the pool of user patterns. Cinematographic rules are applied to retarget a high-definition video for small-screen devices based on the ROIs. According to their user studies the approach produces results comparable to handmade retargeting by experts. However, their approach generalizes poorly to new videos with no user zoom preference data.

As opposed to these works, our method for area of interest identification exploits the interest implicitly shown by the camera users on a particular area. We achieve this by analyzing the location of the cameras and how they are pointed.

2.5. Selection of Optimal Cameras. In [18] the authors propose a system for automatic selection of viewpoint from a set of cameras recording a scene. Their aim is to create one real-time video stream edited according to a set of cinematographic rules based on person tracking (body, head, and hands). Different criteria are used for estimating view suitability such as the tracked person position within the view, relative orientation to the camera based on the estimated direction of movement of the person, detection of skin blobs within the view, and positional relations of the cameras to approximate the action axis rule (also known as 180-degree rule). They also describe methods for video retargeting and viewpoint interpolation by extracting 3D information with a plane-sweeping algorithm. The approach assumes fixed camera positions. In [19] the authors present a method for autonomous viewpoint switching according to perceptual pleasantness, game semantics, viewing device constraints, and user preferences in the context of basketball game multicamera recordings. Context-dependent trade-offs between the introduced concepts of "completeness" (i.e., displaying all relevant information), "closeness" (i.e., displaying details), and "smoothness" (i.e., perceptual and semantic continuity) are used as the basis of a two-way hierarchical view switching approach. A fixed camera setup is used in the work.

2.6. Discussion about Prior Art. Most of the discussed methods for analysis of video are based on content analysis; thus they are computationally expensive. Moreover, exploiting only one or few data modalities may not be sufficient for

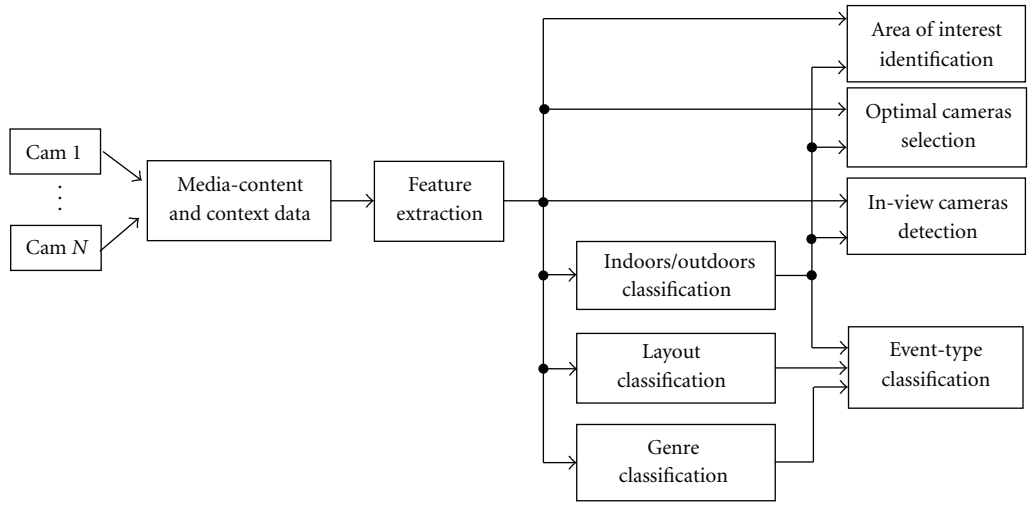

Figure 2: Processing steps for extracting semantics from content data and context data (i.e., auxiliary sensor data captured by the electronic compass, the accelerometer, and the GPS receiver) captured by multiple cameras (denoted in the figure as "Cam") in a public event.

describing the recorded scene in a complete way. Some prior works have jointly analyzed video content, audio content, and text, but the joint use of other types of sensors, such as motion sensors (which nowadays are embedded in most smartphones), together with the more traditional content analysis is still not very popular in the research community. Furthermore, apart from few exceptions, previous works do not consider the availability of media captured by different cameras at the same event.

In contrast to the above works, in this paper we propose to jointly analyze auxiliary sensor data and media content data from multiple capturing devices. Utilizing such auxiliary sensor data allow us to perform operations of low computational cost and to obtain information which would otherwise be hard to extract by means of content analysis only (such as the orientation of a camera).

3. Multimodal Semantics Extraction

In this section we describe the proposed methods for multimodal analysis of user-generated videos. We extract semantic information from the camera locations, from the camera attitude (i.e., the orientation of the camera), and from the recorded media content (video and audio). In particular, we perform the following types of analysis:

(i) classification of the event type;

(ii) identification of the main area of interest;

(iii) detection of the optimal set of cameras to be used in a multicamera video production;

(iv) detection of devices which fall within the field of view of cameras.

In all these analysis methods we assume that the user-generated videos have been captured at the same public event and that they are available, together with the associated auxiliary sensor data (to which we also refer as "context" data), to a computing device performing the actual analysis (e.g., a network server). Figure 2 illustrates the processing steps for the proposed semantics extraction methods. It is important to notice that the data (content and context) captured by different cameras at the same event must be aligned to a common timeline, in order to allow for joint analysis. However, the data alignment is not the focus of this work and therefore we do not elaborate on it here. In the following we first introduce the features which are used by our proposed methods.

3.1. Feature Extraction. We extract several features from different data modalities and we introduce them in the following subsections. We provide reasons for extracting these specific features in the description of the proposed analysis methods.

3.1.1. Video Segmentation and Visual Features. We analyze visual features in combination with audio, compass, and accelerometer features for the task of event genre classification. Obviously, as visual features are extracted from video frames, analyzing all the frames of all videos recorded by multiple cameras at each public event is expensive in terms of computational complexity, other than not necessary. In fact, video content usually contains a lot of temporal redundancy, that is, frames which are temporally close to each other are very similar. This is especially true for user-generated videos, which are usually unedited, that is, they do not contain shot boundaries as each video is usually recorded continuously without pauses. Thus, it is reasonable to consider one frame as representative of a certain number of other nearby frames, and to extract visual features only from such representative frames. In order to overcome the aforementioned issues of computational cost, selection of a subset of the original frames should be performed for each video, where the obtained subset represents the whole video.

There exist different strategies for obtaining such a subset of frames. One of the most common strategies consists of temporally segmenting a video by means of *shot boundary detection* techniques. However, some of the visual features that we extract are derived from the changes in subsequent representative frames, which need to be separated by the same number of frames for all videos. Thus, we have considered another strategy that consists of uniformly sampling the video frames, for example, by selecting one frame every ten seconds of video.

In this work we propose to extract the following *global* visual features for each representative frame: *average brightness*, *dominant color*, *Local Binary Patterns* (LBP) [20], and *color layout* [21]. By considering subsequent representative frames we extract also the following features: *difference of average brightness* and *difference of dominant color*.

Furthermore, we extract local visual features detected by means of Dense Scale-Invariant Feature Transform (DSIFT) in order to compare their performance with respect to the global visual features previously described, in terms of genre classification accuracy. DSIFT is an extension of SIFT [22]. In particular, instead of extracting key points in a sparse way, they are densely extracted from the whole image surface, that is, each frame is divided into blocks and SIFT key points are then extracted from each of such blocks.

3.1.2. Audio Features. By analyzing the audio track of each video recorded at a public event we extract a set of features which are then classified by a Bayesian network. For the audio feature extraction and classification we use the work described in [23].

3.1.3. Compass and Accelerometer Data Features. We extract features by analyzing the data captured by the compass and the accelerometer sensors. From the raw electronic compass data (which represent the horizontal orientations of the camera with respect to magnetic North) captured while recording each video, we extract the following features.

(i) *Average horizontal camera orientation* (φ)—for each video we compute the average of all orientations (given by the compass heading) towards which the camera has been pointed during the video recording activity. The average is computed as the *circular mean*. In particular, φ is expressed as degrees with respect to magnetic North.

(ii) *Camera panning rate*—as the horizontal camera orientation is sampled at a relatively high rate (i.e., 10 Hz), it is possible to automatically detect camera panning movements by analyzing raw compass data (as described in [24]). For each video we compute the panning rate as the ratio between the total number of panning movements and the duration of the recorded video.

From the data captured by the accelerometer during the recording of each video we extract the following features.

(i) *Average vertical camera orientation* (α)—by analyzing the static acceleration on each of the three orthogonal axes of the accelerometer it is possible to determine for each instant the angle by which the device is tilted with respect to the horizontal plane. For each video we compute the average α of such instantaneous vertical orientations.

(ii) *Camera tilting rate*—by analyzing the dynamic distribution of the gravity of Earth g ($\sim 9.81\ \mathrm{m/s^2}$) on the three accelerometer axes it is possible to automatically detect camera tilting movements, as also described in [24]. From the detected tilting movements in each video we derive the camera tilting rate.

3.1.4. GPS Data Features. GPS receivers output different types of data. In this work, we consider only the location information, the measurement time, and the lock status. We analyze these data for obtaining the following features.

(i) *Average GPS location*—we use the GPS receiver embedded in most modern mobile phones for obtaining the instantaneous GPS location of the cameras in terms of coordinate pairs (latitude and longitude). In order to cope with errors in estimating the location, we compute the average of all GPS locations obtained for each camera. In doing this we assume that, while recording videos of the event, the person holding the camera has stayed approximately at the same location. Thus, we obtain the average GPS location of each camera.

(ii) *GPS lock status*—GPS receivers need to be able to communicate with a sufficient number of GPS satellites in order to estimate their location. If this requirement is fulfilled then the GPS receiver is "locked," otherwise its status is "not locked." We check the GPS lock status of all the recording devices and we assign the label "locked" or "not locked" to the feature *GPS lock status* if the majority of the devices are, respectively, locked or not locked.

3.2. Event-Type Classification. For event-type classification we consider the following three aspects of public events: the environment where the event is taking place (we also refer to this as scene classification), the layout of the event, and the event genre. First we classify each of these aspects. The event type is then inferred by simply combining the class labels of these aspects, that is, *indoors* versus *outdoors* for the environment, *stadium* versus *nonstadium* for the layout, and *live music* versus *sport* for the genre. Any combination of these class labels is possible and it represents the final classification of the event type.

In the following we discuss how the extracted multimodal features are used to classify each of these three aspects of public events.

3.2.1. Indoor versus Outdoor Scene. The first analysis step that we perform for inferring the type of an event consists of determining whether the event was held indoors or outdoors. We achieve this in a robust yet simple way, by exploiting sensor-data (instead of the more traditional video content data, as described in Section 2) captured by multiple

cameras recording the event. In particular, we use the data provided by the GPS receivers embedded in camera-enabled mobile phones. It is worth noting that for this analysis any portable device which embeds a GPS receiver can be used and not only camera-enabled devices, as our method does not analyze video content data. In particular we exploit only the information regarding the lock status of the GPS receiver. If the device is in an indoor environment (e.g., inside a building) then it would not be able to "see" a sufficient number of satellites (if not at all) and therefore it will not be locked. Therefore we exploit the GPS lock status for understanding whether the devices are indoors or outdoors. However, there are some situations in which, even if the device is outdoors, due to the presence of surrounding buildings or other tall structures the GPS receiver is not able to receive the signal from a sufficient number of satellites; therefore it will not be locked. In such situations, which is common in practice, an indoors/outdoors classification method which relies solely on one GPS receiver would fail. To overcome this, we exploit the multiuser data availability, that is, we consider the GPS lock status of all the GPS-enabled devices present at the public event. In this way, outlier devices that in an outdoor environment are not able to have GPS-receiver locked (e.g., due to tall structures in their vicinities) are isolated and not taken into account. Thus, if most of the devices are locked, we conclude that the event is held outdoors, otherwise indoors.

The classification between indoor and outdoor scene is used not only by the event-type classifier, but also for other analysis steps that we propose in this work. In fact the identification of an outdoor event enables the following methods which use GPS data: detection of the area of interest, selection of optimal cameras, and detection of in-view cameras.

3.2.2. Layout of the Event. Public happenings are usually held in venues which are specifically designed for allowing people attending the event to enjoy it in an optimal and comfortable way. We refer to the particular structure of such venues as the *layout* of the event. Regarding sport events, the most typical layouts are stadiums (consisting of a central field or stage, which is partly or completely surrounded by the area designated for the people attending the event—for example, for football, rugby, volleyball, and tennis matches), circuit tracks (e.g., for Formula 1 races, motorbike races), and more spatially distributed layouts (e.g., for golf, rally races, bike races, and marathons). For live music events, most often the audience is on one or more sides of the performance stage. In a "proscenium stage," which is the most typical type of stage for music performances, the audience stands or sits only on one side (see Figure 3). However, for big music events, stadiums are the preferred venues as they are usually large enough to contain thousands of spectators.

We propose a method which discriminates two categories of layouts.

(i) *Stadium* layout—those layouts that can be regarded as stadium-like (i.e., where the audience/spectators area has elliptical shape—see Figure 4),

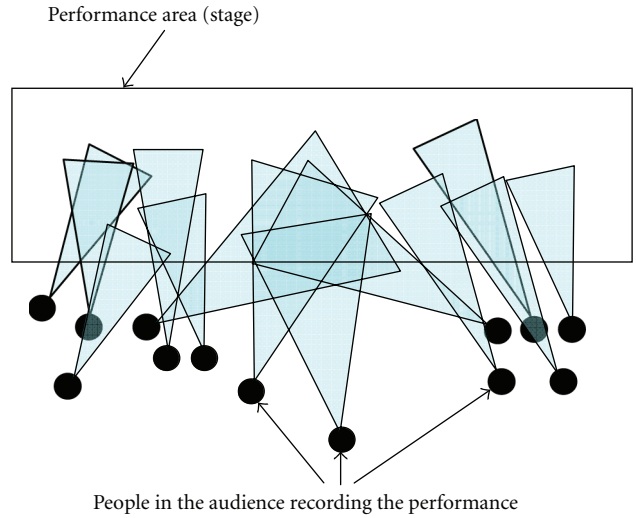

Figure 3: View from the top of a public event (live musical performance) with Nonelliptical layout (nonstadium).

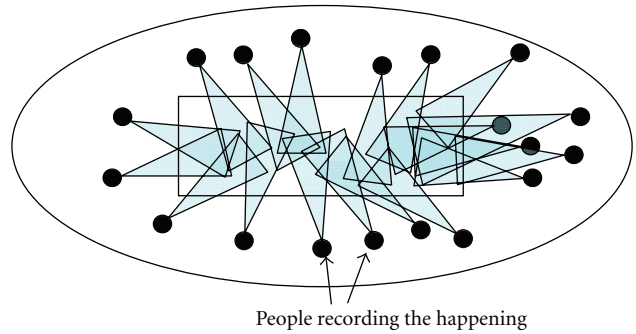

Figure 4: View from the top of a public happening (sport) with elliptical layout (stadium).

(ii) *Nonstadium* layout—those layouts in which the audience/spectators area does not have elliptical shape (e.g., proscenium stages in the case of live music or theater, etc.).

The main idea of the proposed method is to estimate the camera network structure (i.e., how the cameras are spatially distributed and oriented) in order to infer the layout of the event. For this we analyze the locations of the camera users and how they are pointing their camera (i.e., the horizontal camera orientations). Furthermore, we analyze also the tilt angles of the cameras. Location, horizontal orientation, and tilt angle contribute with a different weight to the final classification of the layout. Our method does not perform any video-content analysis to infer the layout of the event which usually requires high computational costs. Figure 5 shows the processing steps required for classifying the layout of an event.

We analyze the GPS position of the cameras to understand whether they are distributed in an elliptical pattern or not. In particular, for each camera we consider its average location throughout the duration of the whole event. If the

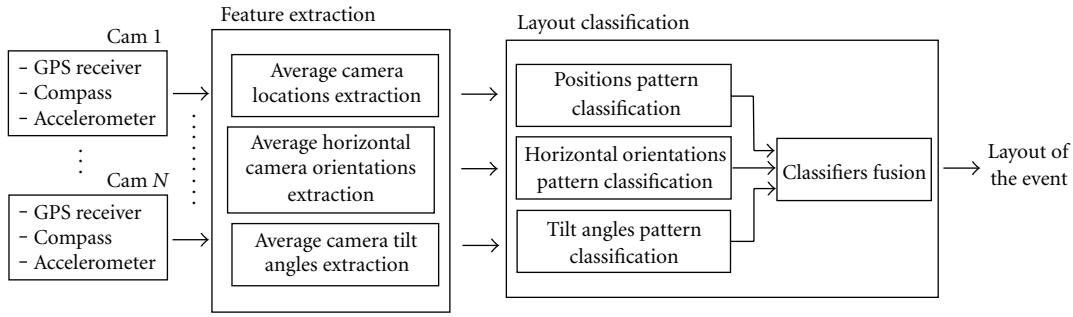

Figure 5: Processing steps for classifying the layout of an event.

camera locations pattern is of elliptical shape, then we assign the nominal value "Elliptical" (numerical value *1*) to it, otherwise "Nonelliptical" (numerical value *0*). In order to classify the camera location pattern, we use an optimization approach that consists of fitting the camera locations to an ellipse and then evaluating the error (i.e., the distance of each camera location from the best-fit ellipse). This can be summarized as shown in Algorithm 1.

Regarding the horizontal orientation information, we consider the average orientation φ that each camera had during the video recording. If the cameras are oriented towards similar orientations, that is, their orientations fall within a predefined and narrow angular range (e.g., 90 degrees), then we assign the nominal value "Directional" (which corresponds to numerical value *0*) to the camera orientations pattern of the event. Otherwise we assign the nominal value "Nondirectional" (numerical value *1*).

Finally, regarding the vertical camera orientation, we consider the average tilt angle α of all the cameras, which represents the most common vertical orientation throughout the recorded event. If the cameras were mostly tilted downwards during the event, we assign a nominal value "stadium" (numerical value *1*) to the tilt angles pattern, otherwise "nonstadium" (numerical value *0*).

For the final classification of the layout, we assign a different weight to the locations pattern loc, the horizontal orientations pattern orient_h and the vertical orientations pattern orient_v. Then we use the numerical values of the patterns for computing a weighted average:

$$\text{layout} = \frac{w_l \cdot \text{loc} + w_h \cdot \text{orient}_h + w_v \cdot \text{orient}_v}{w_l + w_h + w_v}, \quad (1)$$

where w_l, w_h, and w_v are nonnegative weights. Each weight represents the confidence on the discriminative power that each pattern has in the considered layout classification problem. These weights can be obtained through a supervised learning step. However in our case we have assigned the weights empirically after performing extensive experimentations. The final decision on the layout is taken by comparing the weighted average layout with a predefined threshold $\text{Thr}_{\text{layout}}$. If layout is more than $\text{Thr}_{\text{layout}}$ then we classify the event as being held in a *stadium*-like layout, otherwise in a *nonstadium*-like layout.

GPS location information of the cameras is available only for those events held in an outdoor environment. However, if our system detects that the event is held indoors (this information is provided by the indoor versus outdoor scene classification described in Section 3.2.1), the layout classification method will simply not consider location data and it will instead analyze only compass data and accelerometer data.

3.2.3. Event Genre. In video genre classification the most commonly considered genres are *movie*, *news*, *sport*, *music*, *commercials*, and *documentary*, as can be seen also in [2, 8–11]. In this work we consider user-generated videos which have been recorded at a public event. This means that we target specific use cases in which it is likely that a relatively high number of people gather together for attending something of common interest. Thus, we focus on discriminating only between those event genres which comply with this scenario: *sport events* and *live music events*.

We approach the problem of event genre classification by analyzing multiple data modalities collected by multiple cameras present at the event (see Figure 6). In particular we analyze video content data, audio content data, and data from auxiliary sensors (electronic compass, accelerometer). In this way we aim at achieving a robust classification thanks to a more complete description of the scene. As an example, merely applying a simple music occurrence detector to the audio tracks of the recordings and classifying the videos into music or sport genre based on whether the videos contain more music or nonmusic sections, would fail in the following situations: first, in music events people might record things which happened before or after the music show for even longer time than the actual musical performance. Second, many sport events have distinct background music played during breaks and some even during the actual sport activities. Finally, the classification performance easily deteriorates with user-generated real world data—particularly with audio recorded from the audience area using mobile phones, because of the nonprofessional quality of the microphones and because of the background noise originating from the crowd (we have confirmed this experimentally and we give further details on our experiments in the Section 4). Nevertheless, the audio modality contributes significantly to the genre classification task, and its setbacks can be compensated with information from the additional modalities.

(1) *Compute the average location of each camera over the whole time during which that camera has been recording.*
(2) *Apply optimization for fitting the average camera locations to an ellipse, and find the optimal parameters that define the best-fit ellipse.*
(3) *Evaluate the Euclidean distance between each average camera location and the best-fit ellipse. If the average distance is less than a predefined threshold, then we assign the nominal value "Elliptical" to the camera locations pattern of the considered event.*

Algorithm 1: Layout classification based on camera locations.

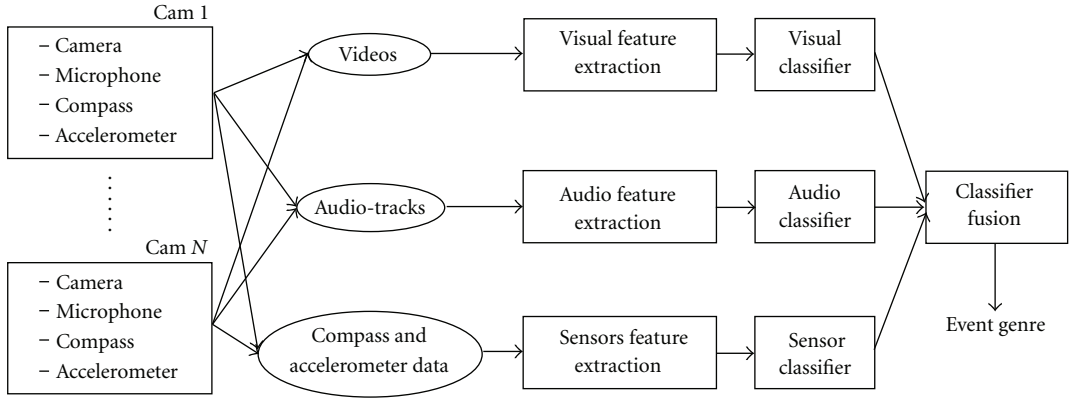

Figure 6: Processing steps for classifying the event genre.

Among the features that we discussed in Section 3.1, for classifying the event genre we use the following ones: *brightness, brightness difference, dominant color, dominant color difference, LBP, color layout, audio features* (described in [23]), *camera panning rate,* and *camera tilting rate*. In Section 4 we compare the classification performances achieved by using these features in different combinations and also when combined with SIFT-based features.

The reasons for choosing this particular set of features are given in what follows. Based on visual and aural inspection of videos belonging to sport and music genres we have noticed big differences in the brightness feature. In fact, live music events are usually held in relatively dark places, whereas sport happenings are characterized by good illumination conditions. Dominant colors are also discriminative for these two event genres, as soccer videos are characterized by green hue, ice-hockey matches by white hue, and so on for other sports, whereas live music events are characterized by many different hues thanks to the frequently changing stage lights, especially by hues such as red, purple, blue; also, in concerts held indoors or at night time another dominant color is usually represented by the black of the scene around the main stage, especially for those cameras which record from longer distances with respect to the stage. Changes in the brightness value and in colors are usually much higher and frequent for live music performances than for sports, due to the stage lights. The texture present in sport videos is usually much more uniform than in live music videos, because of the field (in case of football, ice-hockey, etc.) or tracks (e.g., skiing). Also the color layout was found to be a discriminating feature between the considered genres, as they have different patterns of spatial distributions of colors. For example in football games the green field usually occupies a big portion in the central and bottom parts of the images. Finally, we chose to analyze the camera motion (panning and tilting rate) as it is usually higher when recording sport events (as also stated in [2, 25]).

The actual classification is performed by employing a *late fusion* strategy [26]. As can be seen in Figure 6, each recording device captures data of different modalities, namely, video, audio, compass, and accelerometer data. Feature extraction is performed separately thus obtaining visual, audio, and sensors (compass and accelerometer) feature vectors. The following set of three classifiers is then utilized (one classifier for each data modality).

(i) A Support Vector Machine (SVM) [27] represents the visual classifier, which is used to classify the visual feature vectors.

(ii) A Bayesian network represents the audio classifier, which classifies the audio feature vectors. For this we use the work described in [23]. In particular, we obtain a class label for each temporal segment of predefined length. We then classify the event as the audio class label which occurs most often throughout all videos.

(iii) Another SVM is used to classify the sensors feature vectors.

In Section 4 we give details on how we trained the classifiers. The results of these three classifiers are fused by computing a weighted average where the weights are derived

(1) *Obtain the location and horizontal orientation of each camera which is recording during the considered instant (or temporal segment)—we define such a camera as a recording camera.*
(2) *We derive a linear equation for each recording camera, which represents the direction towards which the camera is pointed. We express the equation in the point-slope form:*
$$y - y_1 = m_1(x - x_1),$$
where the point coordinates (x_1, y_1) are the camera location coordinates, and the slope m_1 is derived from the horizontal orientation.
(3) *For each pair i of recording cameras we solve a system of two such linear equations representing the pointing directions of the considered cameras. By solving each system we find the intersecting point P_i^{int} between the two pointing directions. As a result we obtain a set S_t^{int} of intersections between all the pairs of recording cameras, for the considered instant t.*
(4) *We apply a clustering on all intersecting points, so as to discover the main cluster. In this way we are able to isolate outlier intersections which do not belong to the area of interest. The obtained main cluster represents the instantaneous area of interest A_t of the event.*
(5) *For each instantaneous area of interest we determine its representative point of interest C_t as the centroid of the main cluster. In particular we compute C_t as the average of the intersecting points belonging to the main cluster.*

Algorithm 2: Area of interest identification.

from the classification performance of each single classifier on a test set, that is, each weight represents the confidence of the respective classifier. In the results section we also provide a comparison on the genre classification accuracy achieved by using different combinations of features.

3.3. Area of Interest Identification. In some application scenarios, such as in video content retrieval, it is important to identify the area that attracts the attention of people attending a public event for which videos have been recorded. We propose a novel method for automatically identifying this *area of interest* (we refer to it also as the *AOI*) by analyzing only auxiliary sensor data. Our method is based on the fair assumption that the interest of those people recording videos at the event represents a good indicator of the general interest of all the other attendees, especially when the number of recording cameras is statistically significant. We propose to analyze the way the camera persons are recording at a given time instant t, in order to identify the instantaneous area of interest A_t of the event (see Figure 7). By combining all the instantaneous areas of interest identified throughout the whole duration of the event, we then obtain the main area of interest of the event A_{main}.

In particular our method exploits the availability of camera location and camera horizontal orientation information. Therefore we do not analyze video or audio content, which would require high computational costs.

For each instant t (or temporal segment of predefined length), we perform the steps in Algorithm 2.

The result of applying these steps for every instant (or temporal segment) is a set of instantaneous areas of interest. The main area of interest A_{main} of the event is then derived simply by averaging the coordinates of the intersections forming all instantaneous areas of interest (we use a trimmed mean so as to isolate outlier instantaneous AOIs). We determine the main point of interest C_{main} by computing a trimmed mean of the coordinates of all the instantaneous points of interest.

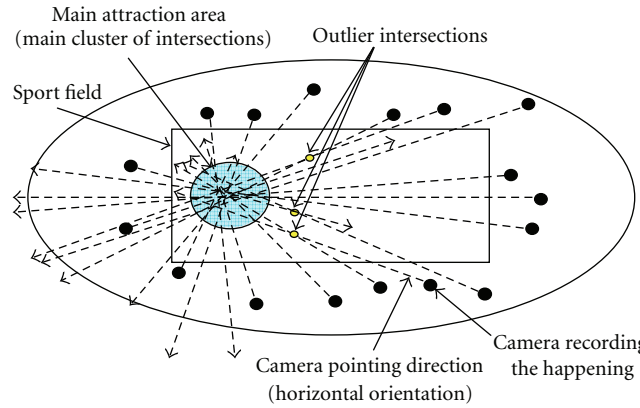

Figure 7: Identification of the area of interest of a public event held in a stadium (view from the top).

3.4. Selection of Optimal Set of Cameras. In multicamera video production (i.e., generation of a mash-up of videos capturing the same scene from multiple cameras) it is important to respect one of the most widely used techniques in filmmaking: the *180-degree rule* [28]. Such a rule is necessary in order not to confuse the viewer of the final video mash-up with regard to the direction of movement of objects within the scene. For example, in the particular case of a football match, the direction of movement of the ball should be consistent when there happens to be a view-switch (i.e., a switch between different cameras) in the video mash-up. In professional video broadcast of football matches this is achieved by placing the cameras only on one side of the football field. This same rule applies for any other type of scene which is recorded by multiple cameras, such as for interviews or live music shows.

Unfortunately, when handling user-generated videos captured by multiple cameras it is not possible to assume that the cameras lied only on one side of the main scene, since users recording the videos can be located anywhere

(1) *Determine the main point of interest P, by applying the method described in Section 3.3.*
(2) *Consider all the lines which intersect the main point of interest, each line having a different slope m and representing a candidate separating line. For each line, count the number of cameras that lie on each of the two sides of the line.*
(3) *Select the candidate separating line which yields the maximum number of cameras on one of its two sides. The selected line represents the optimal separating line, and the optimal set of cameras to be used for generating video mash-ups is made of those cameras on the most populated side of the optimal separating line.*

Algorithm 3: Selection of optimal cameras.

in the stadium or audience area. Therefore there is a need to determine how such cameras were positioned during the event, in order to be able to utilize only those ones which are compatible with the 180-degree rule in the production of a video mash-up. These cameras constitute the *optimal set of cameras*, which is a subset of all cameras recording the event.

We propose a method for automatically determining the optimal set by selecting those cameras which lie on only one of the two sides of the *optimal separating line*. A separating line is an imaginary line which divides the recorded scene into two parts. For example, in Figure 8 a football field and one possible separating line are illustrated. In our method a separating line is determined by two parameters: the point $P(x_P, y_P)$ which is intersected by the line and the slope m of the line. In particular, the intersection point must lie within the recorded scene (i.e., within the area of interest, such as the football field in a football match or the performance stage in a live music show). The optimal separating line is characterized by the optimal slope m_{opt} which yields the maximum number of cameras on one of the two sides of the separating line:

$$y - y_P = m_{\text{opt}}(x - x_P). \tag{2}$$

Our method relies exclusively on the locations of the recording cameras and on a representative point of the main scene. In particular, the GPS locations of all the cameras present at the event are analyzed. We consider as the representative point P of the main scene the *main point of interest* (determined with the method described in Section 3.3). The method can be summarized in the steps shown in Algorithm 3.

3.5. Detection of In-View Cameras. We propose a method for automatically detecting the presence of cameras within the field of view (FOV) of a recording camera, during a public event (see Figure 9). The method is not restricted to detect only camera devices, but it can be used for detecting the presence of any other device for which location information is available, such as GPS-enabled devices. Potential uses of the proposed method are mainly in the field of automatic video mash-up generation and, more in general, in video content retrieval. For example, it would be beneficial to know which specific persons (who hold a GPS-enabled device) are likely to be in the view of the recording camera, or to know whether a camera is recording approximately the same scene as other cameras present at an event.

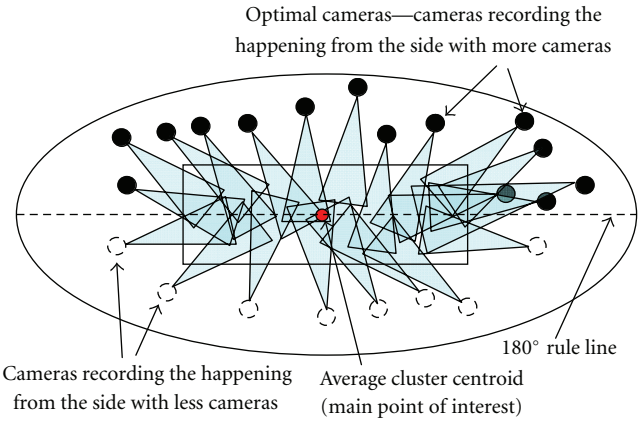

Figure 8: Identification of the 180-degree rule line (the *separating line*) and selection of optimal cameras (view from the top).

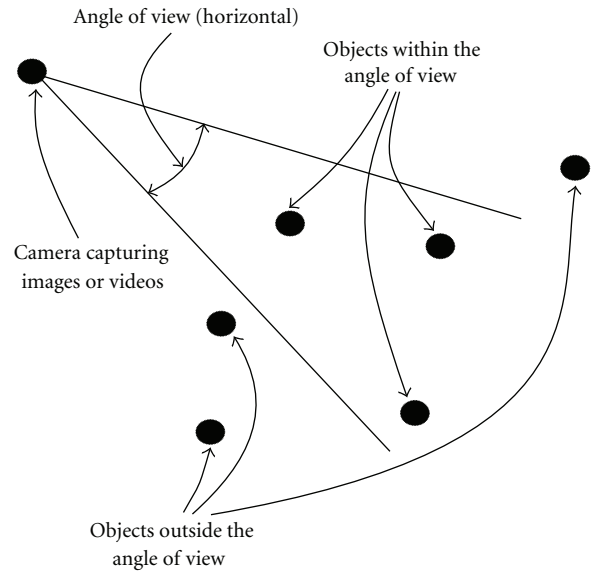

Figure 9: Detection of in-view cameras (view from the top).

For each camera which is recording an event, our method exploits the availability of the location, the pointing orientation, and the field of view of the camera. Furthermore the location of any other GPS-enabled device present at the event is considered. The method consists of the steps described in Algorithm 4.

(1) *Determine the slopes* m_1^{FOV} *and* m_2^{FOV} *of the two lines* l_1^{FOV} *and* l_2^{FOV} *which delimit the field of view of the recording camera, by using the camera pointing angle (horizontal orientation) and the angle of view.*
(2) *Determine the slope* m_i *of each line* l_i *connecting the position* P_c *of the recording camera to the position* P_i *of each other device* i *present at the event.*
(3) *Determine whether each device lies within the determined field of view, by evaluating the slope* m_i *of each line* l_i *with respect to the slopes of the two lines* l_1^{FOV} *and* l_2^{FOV}. *If* m_i *is within the range* $[m_1^{\text{FOV}}, m_2^{\text{FOV}}]$, *then the device* i *is considered to be within the field of view of the recording camera.*

Algorithm 4: Detection of in-view cameras.

4. Results

In this section we evaluate the performance of the proposed methods. As we analyze streams of sensor measurements recorded simultaneously with the video recording, there are no publicly available datasets that already contain such sensor data. In addition, we analyze data captured by multiple devices at the same time and at the same event. Therefore, for our experiments we use test datasets obtained as described in Section 4.1.

4.1. Test Datasets. We used publicly available smart phones and simple dedicated software to enable collection of the sensor data synchronously with video recording. The default sampling rate for each sensor was used, that is, 40 samples/second for the accelerometer, 10 samples/second for the compass, and 1 sample every 5 seconds for the GPS. Each sample is further labeled with its timestamp. 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 sensor measurements (and associated timestamps) as data streams associated with the recorded video.

We used two datasets for our experiments. Dataset 1 contains data (user-generated videos and associated context data) collected at public events of both sport genre and live music genre, held either indoors or outdoors and in stadium or nonstadium layouts. In particular the recorded events were the followings: three football matches held in outdoor stadiums (the number of videos is 54, the total length of all videos is about 720 minutes), two ice hockey matches held in indoor stadiums (71 videos with total length of about 684 minutes), four live music performances held in proscenium stages, from which two were outdoors and two indoors (156 videos for all four events, spanning an overall duration of 890 minutes). The data was collected by multiple users that were attending the events and were sparsely located in the audience (or among the spectators in the case of sport events). Dataset 1 has been used for testing the event-type classification. Dataset 2 contains a subset of the events included in Dataset 1. This second dataset has been used for testing the identification of the area of interest of the event, the selection of optimal cameras, and the detection of in-view cameras. Dataset 2 includes only outdoor events.

Table 1: Event genre classification results by analyzing only audio content.

Event	Ground truth audio class	Automatically extracted audio class
Football match 1	No music	Music
Football match 2	No music	No music
Football match 3	No music	Music
Ice-hockey match 1	No music	Music
Ice-hockey match 2	No Music	Music
Concert 1	Music	Music
Concert 2	Music	Music
Concert 3	Music	No music
Concert 4	Music	Music

In particular, the events belonging to this dataset are three football matches and two live music shows. It is worth noting that the people recording the various public events were not given any specific instructions on the way of recording. On the contrary, they were only asked to record the event as they would normally do when they want to obtain videos for their personal use.

4.2. Classifying the Type of Event. In order to evaluate the event-type classifier, we present our experimental results obtained by classifying each of the events in Dataset 1 according to the following aspects: indoor/outdoor scene, layout, and event genre. These aspects define the final event type.

In particular, regarding the event genre classification, before analyzing multiple data modalities we made some evaluations on using only audio classification for achieving discrimination between *sport* and *live music* event genres. For this, the audio classifier described in [23] has been applied on audio content extracted from the video recordings belonging to Dataset 1. Classification results are given in Table 1. As can be seen, the performance of the audio classifier on user-generated data is not very high even though in our experiments we used high-end phones embedding microphones of higher quality than the most common devices that people use. As an outcome of this preliminary test, we decided to analyze additional data modalities apart from audio content data to achieve event genre classification, as already described in Section 3.2.3.

As we use a supervised classification approach, the three classifiers used for genre classification were firstly trained.

Table 2: Classification of events according to scene, genre, and layout.

Event	Indoor/outdoor scene classification		Event genre classification		Layout classification	
	Ground truth	Proposed method	Ground truth	Proposed method	Ground truth	Proposed method
Football match 1	Outdoor	Outdoor	Sport	Sport	Stadium	Stadium
Football match 2	Outdoor	Outdoor	Sport	Sport	Stadium	Stadium
Football match 3	Outdoor	Outdoor	Sport	Sport	Stadium	Stadium
Ice-hockey match 1	Indoor	Outdoor	Sport	Sport	Stadium	Stadium
Ice-hockey match 2	Indoor	Indoor	Sport	Sport	Stadium	Stadium
Concert 1	Outdoor	Outdoor	Live music	Live music	Nonstadium	Nonstadium
Concert 2	Outdoor	Outdoor	Live music	Live music	Nonstadium	Stadium
Concert 3	Indoor	Indoor	Live music	Live music	Nonstadium	Nonstadium
Concert 4	Indoor	Indoor	Live music	Live music	Nonstadium	Nonstadium
Classification accuracy (%)	—	88.9	—	100	—	88.9

Regarding the *visual classifier*, the training was performed by using the Columbia Consumer Video (CCV) database [29] as the training data, which is a dataset of YouTube videos of different topics, such as several types of sport (soccer, skiing, and ice skating), music performances, and wedding ceremonies. In particular, we selected only those videos which are labeled as sport or music performances.

Regarding the *audio classifier*, as we already mentioned, the work described in [23] was used to classify the audio track of each video as either "music" or "No-music." The Bayesian network has been trained with data captured by mobile phones.

Regarding the *sensor classifier*, there are no publicly available datasets of compass and accelerometer data captured during video recording. Therefore the training dataset is made of sensor data captured by our phones during public happenings. In particular, we considered a different set of phones with respect to those used for testing the classification performance, in order to obtain a training set and a testing set which are as much independent as possible.

The experimental results on classifying scene, layout, and genre by analyzing multiple data modalities are presented in Table 2. Regarding the layout classification for *Concert 2*, we obtained a misclassification (*stadium* instead of *nonstadium*) because the event was held in a big venue in which the camera users happened to be distributed almost in an elliptical way.

Based on the classification accuracies reported in Table 2, the proposed event-type classification method performs well in real-world usage scenarios.

We also performed a comparison on the use of different sets of features for event genre classification. In particular, we analyzed the classification performance for each of the following feature-sets:

(i) *feature-set* S_1: audio features only;

(ii) *feature-set* S_2: sensors features only;

(iii) *feature-set* S_3: only DSIFT (Bag-of-Visual-Words approach);

(iv) *feature-set* S_4: only global visual features;

(v) *feature-set* S_5: combination of audio and sensors features;

(vi) *feature-set* S_6: combination of DSIFT and sensors features;

(vii) *feature-set* S_7: combination of global visual features and sensors features;

(viii) *feature-set* S_8: combination of audio features, DSIFT and sensors features;

(ix) *feature-set* S_9: combination of audio features, global visual features and sensors features. This is the set that we propose to use.

The Bag-of-Visual-Words approach is one of the state-of-the-art methods for classifying images and videos, apart from being used also for detecting objects and salient events. One work in which the video genre is classified using BoVW is the one presented in [8]. The BoVW approach works in two phases:

(1) codebook generation and classifier training phase;

(2) classification phase.

For both phases we densely extract a set of SIFT points from each frame of each video. In the first phase (codebook generation) the points extracted from training videos are clustered into a set of code words using the k-means clustering algorithm. For each representative frame we derive a histogram of code words occurrences and this is achieved by mapping the extracted SIFT points to the obtained code words. The obtained histograms are then used to train a SVM classifier. In the second phase (classification), we consider each representative frame of each video and, by mapping the extracted set of SIFT points to the previously generated code-words, we obtain a histogram of code-words occurrences. Such a histogram represents the feature vector which will be classified by the SVM trained in the first phase.

The results of this comparison are reported in Tables 3 and 4. In particular, in Table 4 we report on the classification results obtained by combining features from different modalities of data. Our proposed approach (feature-set S_9) which uses a combination of audio features, global visual features, and sensors features performs the best in terms of classification accuracy.

Table 3: Performance comparison for the event genre classification task using different feature-sets.

Event	Ground truth event genre	Automatic event genre classification			
		Feature-set S_1 (audio)	Feature-set S_2 (sensors)	Feature-set S_3 (DSIFT)	Feature-set S_4 (global visual features)
Football match 1	Sport	Live music	Sport	Sport	Sport
Football match 2	Sport	Sport	Sport	Sport	Sport
Football match 3	Sport	Live music	Sport	Sport	Sport
Ice-hockey match 1	Sport	Live music	Sport	Live music	Sport
Ice-hockey match 2	Sport	Live music	Sport	Live music	Live music
Concert 1	Live music	Live music	Live music	Live music	Live music
Concert 2	Live music	Live music	Live music	Live music	Sport
Concert 3	Live music	Sport	Live music	Live music	Live music
Concert 4	Live music	Live music	Sport	Live music	Live music
Total accuracy (%)	—	44.4	88.9	77.8	77.8

Table 4: Performance comparison for the event genre classification task using different feature-sets.

Event	Ground truth event genre	Automatic event genre classification				
		Feature-set S_5 (audio, sensors)	Feature-set S_6 (DSIFT, sensors)	Feature-set S_7 (global visual, sensors)	Feature-set S_8 (audio, DSIFT, sensors)	Feature-set S_9 (audio, global visual, sensors)—Proposed set
Football match 1	Sport	Sport	Sport	Sport	Sport	Sport
Football match 2	Sport	Sport	Sport	Sport	Sport	Sport
Football match 3	Sport	Sport	Sport	Sport	Sport	Sport
Ice-hockey match 1	Sport	Sport	Sport	Sport	Live music	Sport
Ice-hockey match 2	Sport	Sport	Sport	Sport	Live music	Sport
Concert 1	Live music	Sport	Sport	Sport	Live music	Live music
Concert 2	Live music	Live music	Live music	Live music	Live music	Live music
Concert 3	Live music	Live music	Live music	Live music	Live music	Live music
Concert 4	Live music	Live music	Live music	Live music	Live music	Live music
Total accuracy (%)	—	88.9	88.9	88.9	77.8	100

Table 5: Experimental results on area of interest identification (AOI) applied on Dataset 2. MSE stands for Mean Square Error.

Event	Number of instantaneous AOIs	Main AOI (identified versus not identified)	MSE of distances from the identified main AOI (meters)
Football match 1	930	Identified	6.8
Football match 2	756	Identified	5.7
Football match 3	549	Identified	31.1
Concert 1	555	Identified	13.2
Concert 2	345	Identified	44.0

4.3. Identifying the Area of Interest. Videos belonging to Dataset 2 have been used for testing the identification of the area of interest. We performed an evaluation by visually estimating where the main area of interest of the whole event is. We plotted the obtained locations of the main area of interest and then visually evaluated whether it has been identified correctly or not. In particular, for sport events we mark the estimated main area of interest as "Identified" if it was located within the football field. For live music events we mark it as "Identified" if it was on the stage (or slightly behind the stage). Table 5 summarizes our experiments on the area of interest identification. In the table, for each recorded event, we report the total number of instantaneous areas of interest (i.e., the number of analyzed temporal segments), the identification of the main AOI, and the *Mean Square Error* (MSE) of the distances from each camera to the identified main AOI. We are able to identify the main AOI in all the events of Dataset 2. Furthermore we obtain different accuracies for the estimated distances between cameras and area of interest. In particular, for *Concert 2* we obtained the highest MSE value, which is due to an identification of the main AOI behind the performance stage (which represents the ground truth main AOI) and due to inaccuracies in the GPS measurements.

4.4. Selecting the Optimal Set of Cameras. Tests on the selection of the optimal cameras according to the 180-degree rule have been carried out on Dataset 2. The method that we proposed for determining the optimal cameras relies on the correct identification of the main area of interest of the event; in particular it uses the center of the main AOI and considers

Table 6: Experimental results on identifying the optimal cameras. *P* stands for *Precision*, *R* for *Recall,* and *F* for *F-measure*.

Event	All cameras (index)	Ground truth sets of optimal cameras	Automatically identified optimal set of cameras	*P*	*R*	*F*
Football match 1	[1, 2, 3, 4, 5, 6]	[1, 2, 3], [2, 3, 4, 5], [5, 6]	[2, 3, 4, 5, 6]	0.8	1.0	0.89
Football match 2	[1, 2, 3, 4, 5, 6]	[1, 2, 3, 4], [4, 5, 6]	[1, 2, 3, 4]	1.0	1.0	1.0
Football match 3	[1, 2, 3, 4, 5]	[1, 2], [2, 3], [4, 5]	[1, 2, 3]	0.75	1.0	0.86
Concert 1	[1, 2, 3, 4, 5, 6, 7, 8, 9]	[1, 2, 3, 4, 5, 6, 7, 8, 9]	[1, 2, 3, 4, 5, 6, 7, 8, 9]	1.0	1.0	1.0
Concert 2	[1, 2, 3, 4, 5, 6, 7, 8,]	[1, 2, 3, 4, 5, 6, 7, 8]	[1, 2, 3, 4, 5, 6, 7, 8]	1.0	1.0	1.0
Average over all events	—	—	—	0.91	1.0	0.95

Table 7: Experimental results for detecting in-view cameras. *P* stands for *Precision*, *R* for *Recall,* and *F* for *F-measure*.

Event	*P*	*R*	*F*
Football match 1	0.86	0.71	0.77
Football match 2	0.78	0.58	0.67
Football match 3	1.0	0.78	0.88
Concert 1	0.74	0.69	0.71
Concert 2	0.76	0.81	0.78
Average over all events	0.83	0.71	0.77

this as the point intersected by the separating line. Regarding the ground truth, there could be more than one optimal set of cameras, and this was taken into account in our experiments. The experimental results are reported in Table 6.

We use the following measures for evaluating the performance of the selection method:

(i) precision (*P*)—fraction of the automatically selected cameras which belong to one of the ground truth sets of optimal cameras;

(ii) recall (*R*)—fraction of the optimal cameras belonging to one of the ground truth optimal sets which are correctly selected by our method;

(iii) balanced *F*-measure (*F*)—it is computed as the harmonic mean of the precision and recall.

As can be seen in Table 6, for *Football match 1* and *Football match 3* our method has introduced one additional camera with respect to one of the ground truth optimal sets. This error was caused by the inaccuracies in the GPS data measurements. Regarding *Concert 1* and *Concert 2*, as the shows were held in proscenium stages and all the recording cameras were located in front of the stage, the ground truth optimal sets include all the cameras. The proposed method correctly identified these optimal sets.

4.5. Detecting In-View Cameras. We have tested the detection of cameras which are within the field of view of other cameras by using Dataset 2. Table 7 summarizes the experimental results. For evaluating the performance of the detection method we use similar measures as for the selection of optimal cameras.

In particular, (i) precision (*P*)—fraction of the detected cameras which are indeed in the field of view of other cameras; (ii) recall (*R*)—fraction of the true in-view cameras which are detected correctly; (iii) balanced *F*-measure (*F*)—it is computed as the harmonic mean of the precision and recall.

5. Discussion

User-generated content has seen a tremendous growth during the latest years [30] and the analysis of such content is becoming an important research problem. In this work we show that context data from multiple user-generated videos can provide important information about the environment in which they were recorded. This information can subsequently be exploited by various other applications (such as video retrieval, summarization, and mash-up creation).

One of the main contributions of this work is the exploitation of multiple modalities of data for analyzing user-generated content. The auxiliary sensor modalities not only allow for precise information about the location and the orientation of the recording device but also their processing involves much less computations than traditional content analysis methods. For example, all the auxiliary sensors that we use in this work produce less than 200 samples per second, whereas one second of HD video content at 25 frames per second contains 23 million pixels.

In this work we used GPS data for indoor/outdoor scene classification, for identification of event layout and area of interest, for selection of optimal cameras, and for detection of in-view cameras. The GPS is usually available only for public events held outdoors. However, if an indoor positioning system is available then our methods can be easily extended to indoor events. As GPS location information is affected by errors originated from several sources, it is worth discussing the effects of such errors on the methods that we proposed in this paper. In a recent paper [31] the authors claim that the average location error experienced on modern mobile phones vary between 8 and 12 meters. In the work described in [32] (from 2011) mobile phones are considered and GPS errors are reported to be between 0 and 5 meters. In particular, different models of modern smartphones are tested for estimating the GPS inaccuracies. For one of such models, 97% of the measurements were found to be affected by errors within 5 meters. As we already mentioned, in order to cope with GPS inaccuracies and especially with outlier location measurements, we capture the location information

multiple times for each camera and then we compute a trimmed mean of such measurements.

Regarding the method that we proposed for indoor/outdoor scene classification, inaccuracies in the GPS location information do not affect the performance of our algorithm. However, if the recorded event is held outdoors and most of the GPS receivers are not locked then our method would provide wrong information. This situation might happen when the event area is small and it is surrounded by tall buildings or other structures. In stadium-like venues there are usually no buildings which are too close to the event area, and we have experimentally verified that the structures which constitute the spectator sections do not represent major problems in terms of direct line of sight.

Regarding the identification of the event layout, as typical positioning errors in mobile phones are within 5 or 10 meters and stadiums have much larger dimensions, such errors do not have big effects on the estimation of the layout, that is, it is still possible to determine if the cameras are distributed in an elliptical way by using our approach based on curve fitting. We have proven this experimentally in our tests in which all the stadium-like venues were correctly identified.

The proposed method for identifying the area of interest is more sensitive to GPS positioning inaccuracies. However, we do not aim at precisely determining the exact position where the interest point (or focus point) is located; instead we are interested in identifying a wider area that can give indicative information about where the show (sport match or music performance) is located within the whole event area. Therefore, slightly inaccurate location measurements (as those previously discussed) do not interfere with this goal.

The selection of optimal cameras relies on the identification of the area of interest and on the position of each camera with respect to such area. The performance of this method could be impaired by inaccuracies in the location information. In fact, because of such errors, a camera which is in reality on one side with respect to the 180 degree line can be erroneously detected as being on the other side. We have experienced this in our tests (Table 6), in which the automatically selected optimal cameras not always completely corresponded to the ground truth optimal cameras. Finally, regarding the proposed method for detecting cameras which fall within the field of view of other cameras, location inaccuracies in both the recording camera and the target cameras could affect the results. In fact, if a target camera is close to the border of the field of view of the recording camera, even small location errors in either the target or the recording camera can affect the detection accuracy. Another case in which small GPS inaccuracies would produce incorrect detection results is when the recording camera and the target camera are close to each other.

6. Conclusions

In this work we propose a set of methods for automatically extracting semantic information about public happenings such as sport and live music events. The methods rely on the analysis of user-generated videos recorded at those events by multiple recording devices. In particular, we extract information about the recorded scene by taking into account the locations of the cameras and other contextual information of the recording activity. Auxiliary sensor data, together with video and audio content data, is analyzed for determining the type of event being recorded. In particular, we are able to identify the layout of the event, the event genre, and whether the event is held indoors or outdoors. Furthermore, we have proposed algorithms for identifying the area of interest of an event and for automatically selecting the optimal set of cameras to be used for a multicamera video production, according to the 180-degree rule which is a widely used technique in filmmaking. Finally a method for detecting devices which are within the field of view of cameras was described. We performed experiments for evaluating the proposed algorithms on real test data. In particular we obtained the following classification accuracies for, respectively, scene, genre, and layout: 88.9%, 100%, and 88.9%. The main area of interest has been identified in all the test cases. By using the identified main areas of interest, we were able to select the optimal cameras with an average F-measure of 0.95. Finally, for the detection of in-view cameras we obtained an average F-measure of 0.77. Thus, our experimental results show that the proposed methods perform well in several real public events.

References

[1] R. Oami, A. B. Benitez, S. F. Chang, and N. Dimitrova, "Understanding and Modeling User Interests in Consumer Videos," in *IEEE International Conference on Multimedia and Expo*, pp. 1475–1478, Taipei, Taiwan, 2004.

[2] M. Sugano, T. Yamada, S. Sakazawa, and S. Hangai, "Genre Classification Method for Home Videos," in *IEEE International Workshop on Signal Processing*, pp. 1–5, Rio de Janeiro, Brazil, 2009.

[3] D. Brezeale and D. J. Cook, "Automatic video classification: a survey of the literature," *IEEE Transactions on Systems, Man and Cybernetics C*, vol. 38, no. 3, pp. 416–430, 2008.

[4] N. Serrano, A. Savakis, and J. Luo, "A computationally efficient approach to indoor/outdoor scene classification," in *16th IEEE International Conference on Pattern Recognition*, pp. 146–149, Quebec City, Canada, 2002.

[5] U. Lipowezky and I. Vol, "Indoor-outdoor detector for mobile phone cameras using gentle boosting," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 31–38, San Francisco, Calif, USA, 2010.

[6] M. Szummer and R. W. Picard, "Indoor-outdoor image classification," in *IEEE International Workshop on Content-Based Access of Image and Video Database*, pp. 42–51, Bombay, India, 1998.

[7] A. Payne and S. Singh, "Indoor vs. outdoor scene classification in digital photographs," *Pattern Recognition*, vol. 38, no. 10, pp. 1533–1545, 2005.

[8] N. Zhang and L. Guan, "An efficient framework on large-scale video genre classification," in *IEEE International Workshop on Multimedia Signal Processing*, pp. 481–486, Saint-Malo, France, 2010.

[9] X. Yuan, W. Lai, T. Mei, X. S. Hua, X. Q. Wu, and S. Li, "Automatic video genre categorization using hierarchical SVM," in *IEEE International Conference on Image Processing*, pp. 2905–2908, Atlanta, Ga, USA, 2006.

[10] J. Xinghao, S. Tanfeng, and C. Bin, "A novel video content classification algorithm based on combined visual features model," in *2nd International Congress on Image and Signal Processing (CISP '09)*, October 2009.

[11] M. Montagnuolo and A. Messina, "Multimodal genre analysis applied to digital television archives," in *19th International Conference on Database and Expert Systems Applications (DEXA '08)*, pp. 130–134, Turin, Italy, September 2008.

[12] A. Feryanto and I. Supriana, "Location recognition using detected objects in an image," in *International Conference on Electrical Engineering and Informatics*, pp. 1–4, Ban-dung, Indonesia, 2011.

[13] G. Schroth, R. Huitl, D. Chen, M. Abu-Alqumsan, A. Al-Nuaimi, and E. Steinbach, "Mobile visual location recognition," *IEEE Signal Processing Magazine*, vol. 28, no. 4, pp. 77–89, 2011.

[14] K. Tieu, G. Dalley, and W. E. L. Grimson, "Inference of non-overlapping camera network topology by measuring statistical dependence," in *10th IEEE International Conference on Computer Vision*, vol. 2, pp. 1842–1849, Beijing, China, 2005.

[15] T. Thummanuntawat, W. Kumwilaisak, and J. Chinrungrueng, "Automatic region of interest detection in multi-view video," in *International Conference on Electrical Engineering/Electronics Computer Telecommunications and Information Technology (ECTI-CON '10)*, pp. 889–893, Chiang Mai, Thailand, May 2010.

[16] J. B. Hayet, T. Mathes, J. Czyz, J. Piater, J. Verly, and B. Macq, "A modular multi-camera framework for team sports tracking," in *IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS '05)*, pp. 493–498, Como, Italy, September 2005.

[17] A. Carlier, V. Charvillat, W. T. Ooi, R. Grigoras, and G. Morin, "Crowdsourced automatic zoom and scroll for video retargeting," in *18th ACM International Conference on Multimedia ACM Multimedia (MM '10)*, pp. 201–210, Firenze, Italy, October 2010.

[18] P. Doubek, I. Geys, T. Svoboda, and L. Van Gool, "Cinematographic rules applied to a camera network," in *5th Workshop on Omnidirectional Vision, Camera Networks and Non-Classical Cameras*, pp. 17–29, Prague, Czech Republic, 2004.

[19] F. Chen and C. DeVleeschouwer, "Personalized production of basketball videos from multi-sensored data under limited display resolution," *Elsevier Journal of Computer Vision and Image Understanding*, vol. 114, no. 6, pp. 667–680, 2010.

[20] T. Ojala, M. Pietikainen, and D. Harwood, "Performance evaluation of texture measures with classification based on Kullback discrimination of distributions," in *12th IAPR International Conference on Pattern Recognition*, vol. 1, pp. 582–585, Jerusalem, Palestine, 1994.

[21] MPEG-7, "ISO/IEC 15938, Multimedia Content Description Interface," http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=34228.

[22] D. G. Lowe, "Object recognition from local scale-invariant features," in *IEEE International Conference on Computer Vision*, vol. 2, pp. 1150–1157, Corfu, Greece, 1999.

[23] T. Lahti, *On low complexity techniques for automatic speech recognition and automatic audio content analysis*, Doctoral thesis, Tampere University of Technology, 2008.

[24] F. Cricri, K. Dabov, I. D. D. Curcio, S. Mate, and M. Gabbouj, "Multimodal Event Detection in User Generated Videos," in *IEEE International Symposium on Multimedia*, pp. 263–270, Dana Point, Calif, USA, December 2011.

[25] V. Kobla, D. DeMenthon, and D. Doermann, "Identification of sports videos using replays, text, and camera motion features," in *Storage and Retrieval for Media Databases*, vol. 3972 of *Proceedings of SPIE*, pp. 332–343, 2000.

[26] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders, "Early versus late fusion in semantic video analysis," in *ACM International Conference on Multimedia*, pp. 399–402, Singapore, 2005.

[27] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, NY, USA, 1995.

[28] B. Foss, *Filmmaking: Narrative and Structural Techniques*, Silman James Press, Los Angeles, Calif, USA.

[29] Y. G. Jiang, G. Ye, S. F. Chang, D. Ellis, and A. C. Loui, "Consumer video understanding: A benchmark database and an evaluation of human and machine performance," in *1st ACM International Conference on Multimedia Retrieval (ICMR '11)*, Trento, Italy, April 2011.

[30] M. Cha, H. Kwak, P. Rodriguez, Y. Y. Ahn, and S. Moon, "Analyzing the video popularity characteristics of large-scale user generated content systems," *IEEE/ACM Transactions on Networking*, vol. 17, no. 5, pp. 1357–1370, 2009.

[31] Y. Odaka, S. Takano, Y. In, M. Higuchi, and H. Murakami, "The evaluation of the error characteristics of multiple GPS terminals," in *Recent Researches in Circuits, Systems, Control and Signals*, pp. 13–21, 2011.

[32] T. Menard, J. Miller, M. Nowak, and D. Norris, "Comparing the GPS capabilities of the Samsung Galaxy S, Motorola Droid X, and the Apple iPhone for Vehicle Tracking Using FreeSim_Mobile," in *14th IEEE International Conference on Intelligent Transportation Systems*, pp. 985–990, Washington, DC, USA, 2011.