Statistical Motion Information Extraction and Representation for Semantic Video Analysis

Georgios Th. Papadopoulos, *Student Member, IEEE,* Alexia Briassouli, Vasileios Mezaris, *Member, IEEE,* Ioannis Kompatsiaris, *Member, IEEE,* and Michael G. Strintzis, *Fellow, IEEE*

***Abstract*—In this paper, an approach to semantic video analysis that is based on the statistical processing and representation of the motion signal is presented. Overall, the examined video is temporally segmented into shots and for every resulting shot appropriate motion features are extracted; using these, Hidden Markov Models (HMMs) are employed for performing the association of each shot with one of the semantic classes that are of interest. The novel contributions of this work lie in the areas of motion information processing and representation. Regarding the motion information processing, the kurtosis of the optical flow motion estimates is calculated for identifying which motion values originate from true motion rather than measurement noise. Additionally, unlike the majority of the approaches of the relevant literature that are mainly limited to global- or camera-level motion representations, a new representation for providing local-level motion information to HMMs is also presented. It focuses only on the pixels where true motion is observed. For the selected pixels, energy distribution-related information, as well as a complementary set of features that highlight particular spatial attributes of the motion signal, are extracted. Experimental results, as well as comparative evaluation, from the application of the proposed approach in the domains of tennis, news and volleyball broadcast video, and human action video demonstrate the efficiency of the proposed method.**

***Index Terms*—Semantic video analysis, motion representation, Hidden Markov Models (HMMs), kurtosis.**

I. Introduction

THE rapid progress in hardware technology has led to an enormous increase of the total amount of video content generated every day and of the available means for distributing it. Additionally, large-scale video archives are now accessible by the average user more than ever, while in many cases access to video archives is part of users' every day activities at either personal or professional level. To this end, new needs and issues arise regarding the efficient and skillful manipulation of video content. This has triggered intense research efforts towards the development of sophisticated and user-friendly systems for the indexing, search and retrieval of video sequences [1].

More recently, the fundamental principle of simulating the way that humans perceive and process the visual information and incorporating such models into video manipulation techniques has been widely adopted. These approaches shift video analysis methods towards a semantic level, thus attempting to bridge the so called *semantic gap* [2]. A wide variety of semantic video analysis approaches have been proposed. In [3], an ontology framework, making use of explicitly defined axioms, facts and rules, is presented for detecting events in video sequences. It is based on the notion that complex events are constructed from simpler ones by operations such as sequencing, iteration and alternation. A large-scale concept ontology for multimedia (LSCOM) is designed in [4] to simultaneously cover a large semantic space and increase observability in diverse broadcast news video data sets. In [5], Support Vector Machines (SVMs), which perform on top of specific feature detectors, are employed for detecting semantically meaningful events in broadcast video of multiple field sports. Additionally, in [6], Bayesian Networks (BNs) are employed for detecting concepts of a lexicon using cues derived from audio, visual and text features. Among the various Machine Learning (ML) techniques, it is Hidden Markov Models (HMMs) [7] that have been used most extensively for video analysis tasks, due to their suitability for modeling pattern recognition problems that exhibit an inherent temporality. In [8], a HMM-based system is developed for extracting highlights from baseball game videos. An approach that supports the detection of events such as 'foul' and 'shot at the basket' in basketball videos is presented in [9].

A prerequisite for the application of any semantic video analysis technique is the compact, appropriate for the analysis task at hand and the adopted analysis methodology, representation of the content low-level properties, such as color, motion, etc. In video analysis, the focus is on motion representation, since the motion signal bears a significant portion of the semantic information that is present in a video sequence. To this end, a series of approaches for the extraction and representation of discriminative motion-based features from the video stream have been proposed [10]. Motion activity features of video segments are utilized for realizing semantic characterization of video content in [11], [12]. Camera-level motion representations are proposed in [13], [14], for performing semantic video annotation. Leonardi et al. utilize motion indices like camera operations and the presence of shot cuts for realizing semantic video indexing [15]. Additionally, the notion of 'motion texture' is introduced in [16] for modeling the motion patterns of a video clip, while Adams et al. use the attributes of motion and shot length to define and compute the so called 'tempo' measure in order to detect

Georgios Th. Papadopoulos and Michael G. Strintzis are with the Electrical and Computer Engineering Department of Aristotle University of Thessaloniki, Greece and CERTH/Informatics and Telematics Institute, Greece (e-mail: papad@iti.gr, strintzi@eng.auth.gr).

Alexia Briassouli, Vasileios Mezaris and Ioannis Kompatsiaris are with the CERTH/Informatics and Telematics Institute, Greece (e-mail: abria@iti.gr, bmezaris@iti.gr, ikom@iti.gr).

particular semantic events in movie videos [17]. Moving to a finer level of granularity, Dagtas et al. [18] use a semi-manual object tracking algorithm for estimating the trajectory of the foreground moving objects and subsequently utilize this information for detecting the semantic events of interest. Roach et al. [19] utilize a quantitative measure of the rate of change in the motion of the foreground objects along with a simple parametric camera motion model for identifying the genre of videos. Moreover, an entropy-based criterion is proposed in [20] to characterize the pattern and intensity of object motion in a video sequence as a function of time.

In addition to the motion representations that have been proposed for semantic video analysis in the general case, a series of well-performing domain-specific approaches, i.e. approaches that exploit specific facts and characteristics of the motion signal that are only present in the examined domain, have also been proposed for tasks like sports [21], [22], [23] and news [24], [25] video analysis, and human action recognition [26], [27], [28]. Other elaborate approaches for motion-based recognition that employ complex statistical models include the extraction of temporal textures [29], Gibbs [30] and Gaussian [31] modeling, and more general spatio-temporal patterns [32], [33].

Regarding more specific motion representations for use together with HMMs, a plurality of approaches have also been proposed. In [34], the dominant motion over the entire image field of view is used for detecting semantic events in rugby broadcast videos. In [35], the motion energy redistribution is calculated for every frame and subsequently a set of motion filters are employed for estimating the frame dominant motion, in an attempt to detect semantic events in various sports videos. Huang et al. consider the first four dominant motion vectors and their appearance frequencies, along with the mean and the standard deviation of motion vectors in the frame, for performing scene classification [36]. Additionally, global-level cinematic features of motion, namely the average motion magnitude, the motion entropy, the dominant motion direction and camera pan/tilt/zoom factors, are used for identifying the genre of sports video in [37]. In [38], the median of the frames' mean motion magnitude values is estimated for every GOP, for realizing video genre classification. Moreover, Gibert et al. estimate the principal motion direction of every frame [39], while Xie et al. calculate the motion intensity at frame level [40], for realizing sport video classification and structural analysis of soccer video, respectively. Although significant research efforts have been devoted for developing generic systems for HMM-based semantic video analysis, the majority of the proposed motion representations is mainly limited to global- or camera-level motion processing approaches and the potential of analyzing the motion signal at local-level has not been sufficiently investigated.

In this paper, an approach to semantic video analysis, which is based on the statistical processing and representation of the motion signal, is presented. Under the proposed approach, the examined video is segmented into shots and motion features are extracted for each estimated shot. These features serve as input to HMMs, which perform the association of each shot with one of the semantic classes that are of interest in a possible application case. The novel contributions of this work lie in the areas of motion information processing and representation. In particular, higher order statistics, namely the kurtosis, of the optical flow motion estimates are calculated for identifying which motion values originate from true motion rather than measurement noise, resulting in the robust estimation of activity areas over a series of frames. This is motivated by the fact that higher-order statistics (including the kurtosis) become zero for Gaussian (or nearly Gaussian) data [41], and can therefore effectively detect outlying signals. In the literature, the kurtosis has been used extensively to separate signals from noise, in blind source separation, equalization, face recognition, system identification [42], [43], [44]. In this work, the kurtosis of the inter-frame illumination changes is shown to produce a more robust estimate of pixel activity than directly considering the optical flow estimates for extracting activity areas via e.g. thresholding. Additionally, unlike the majority of the approaches of the relevant literature that are mainly limited to global- or camera-level motion representations, a new representation for providing local-level motion information to HMMs is presented. It focuses only on the pixels that are characterized as active in the corresponding activity area mask, i.e. the pixels where true motion is observed. For the selected pixels, energy distribution-related information, as well as a complementary set of features that highlight particular spatial attributes of the motion signal, are extracted. As will be seen by the experimental evaluation of the proposed approach, the combination of energy distribution-related information and spatial attributes of the motion signal efficiently captures the semantics present in the visual medium.

The paper is organized as follows: Section II discusses the statistical analysis of the motion signal. The proposed motion feature representation is detailed in Section III. Section IV outlines how HMMs are employed for performing motion-based classification. Experimental results from the application of the proposed approach in various domains, as well as comparative evaluation with other approaches of the literature, are presented in Section V, and conclusions are drawn in Section VI.

II. Statistical Motion Analysis

Most of the previously proposed elaborate approaches for statistical processing of the motion signal (e.g. [29], [45], [31]) are limited to constrained application environments. In this section, a statistical analysis approach is presented for analyzing the motion in various kinds of video. In particular, the proposed method aims to extract reliable information about the activity that is present within a video scene, by estimating the kurtosis of each pixel's activity in order to localize the pixels where true motion is observed.

The motivation for the use of the kurtosis to localize active pixels is the fact that it has been shown to be a robust detector of outliers in Gaussian noise [46], as it is asymptotically insensitive to it. This property has led to its use in numerous applications. In blind source separation, the kurtosis is used to suppress the noise present in each separate source component [44], as its value is maximized when the data is the actual

signal (and not interference noise). The kurtosis has been shown to be a reliable measure of outliers in non-Gaussian noise as well [47], [48], so its use is extended to applications where the noise is not strictly Gaussian. For example, it has been used for face recognition, where noisy components are separated from non-noisy ones by maximizing the kurtosis of the latter [43].

In this work, the kurtosis is used to localize active and static pixels in a video sequence, serving as a measure of each pixel's activity. The kurtosis is shown to have low values when a pixel is static, and higher values in active pixels. Under the proposed approach, the pixel activity is measured using the motion energy estimates described in Section II-A, as they provide meaningful information about the amount of activity taking place [35], [49]. The motion energy for static pixels originates from measurement noise, which is usually modeled by a Gaussian distribution. The motion energy of active pixels will be significantly higher, and an outlier to the randomly distributed measurement noise. Thus, the kurtosis of the active pixels' motion energy is expected to be higher than that of the static pixels.

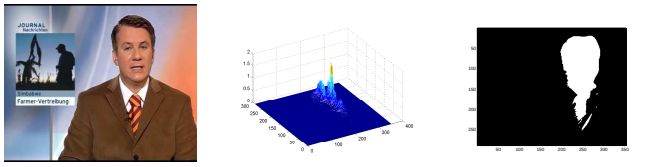
A. Motion Analysis Pre-Processing

The examined video sequence is initially segmented into shots, which constitute the elementary image sequences of video. For shot detection, the algorithm of [50] is used, mainly due to its low computational complexity. This results in a set of shots, denoted by $S = \{s_i,\ i = 1, ...I\}$; under the proposed approach each shot will be associated with one of the supported semantic classes, denoted by $E = \{e_j,\ j = 1, ..J\}$, on the basis of its semantic contents. After shot segmentation, each shot s_i is further divided into a set of sequential non-overlapping time intervals of equal duration, denoted by $W_i = \{w_{ir},\ r = 1, ...R_i\}$, starting from the first frame. The duration of each interval, i.e. the length of the selected time window, is set equal to TW. For every time interval w_{ir}, an individual observation vector will be estimated for representing its motion information, to support shot-class association.

In parallel to temporal video segmentation, a dense motion field is estimated for every frame. The optical flow estimation algorithm of [51] was used for computing this dense motion field, since satisfactory results can be obtained by its application in a variety of motion estimation cases. From the computed dense motion field a corresponding motion energy field is calculated, according to the following equation:

$$M(x, y, t) = \|\overrightarrow{V}(x, y, t)\| \tag{1}$$

where $\overrightarrow{V}(x, y, t)$ is the estimated dense motion field, $\|.\|$ denotes the norm of a vector, and $M(x, y, t)$ is the resulting motion energy field. Variables x, y get values in the ranges $[1, V_{dim}]$ and $[1, H_{dim}]$ respectively, where V_{dim} and H_{dim} are the motion field vertical and horizontal dimensions (same as the corresponding frame dimensions in pixels), whereas variable t denotes the temporal order of the frames. The choice of transforming the motion vector field to an energy field is based on the observation that the latter often provides more appropriate information for motion-based recognition problems [49], [35].



First frame of sequence | Kurtosis field | Activity area mask

Fig. 1. Example of kurtosis field and activity area mask computation for a news broadcast video sequence.

B. Kurtosis Field Estimation and Activity Area Extraction

The motion energy estimates, $M(x, y, t)$, at each pixel represent changes in illumination that originate either from measurement noise, or from pixel displacement (true motion) and measurement noise. This can be expressed as the following hypotheses:

$$\begin{aligned} H_0 : M^0(x, y, t) &= n(x, y, t) \\ H_1 : M^1(x, y, t) &= o(x, y, t) + n(x, y, t), \end{aligned} \tag{2}$$

where $o(x, y, t)$ represents the noiseless motion energy field and $n(x, y, t)$ additive noise. There is no prior knowledge about the statistical distribution of measurement noise, however the standard assumption in the literature is that it is independent from pixel to pixel, and follows a Gaussian distribution [52]. This leads to the detection of which velocity estimates correspond to a pixel that is actually moving by simply examining the non-gaussianity of the data [53]. The classical measure of a random variable's non-gaussianity is its kurtosis, defined by:

$$kurtosis(\psi) = E[\psi^4] - 3(E[\psi^2])^2, \tag{3}$$

where ψ is a random variable. The kurtosis value for Gaussian data is zero.

Since the measurement noise is approximated by a Gaussian distribution, the kurtosis of a static pixel's illumination changes (corresponding to H_0 in Eq. (2)) over a series of frames will also be Gaussian. Consequently, the pixels that undergo true motion can be located by estimating their kurtosis. For this purpose, the kurtosis of every pixel's motion energy estimates $M(x, y, t)$ is calculated over a time interval w_{ir}, according to the following equation:

$$K_{ir}(x, y) = E[M(x, y, t)^4] - 3(E[M(x, y, t)^2])^2, \tag{4}$$

where $K_{ir}(x, y)$ is the estimated kurtosis field and the expectations $E[\cdot]$ are approximated by the corresponding arithmetic means. When a pixel's illumination changes follow a precisely Gaussian distribution, $K_{ir}(x, y)$ will be equal to zero. It should be noted that, even when the unknown noise in the motion estimates deviates from the Gaussian model, the kurtosis remains appropriate for finding the active pixels. This is because their values are outliers, compared to the measurement noise values, and in [48] it is proven that the kurtosis is a robust, locally optimum test statistic for the detection of outliers, even in the presence of non-Gaussian noise.

Following the estimation of the kurtosis field, the distinction between 'active' and 'static' pixels can be made by simple

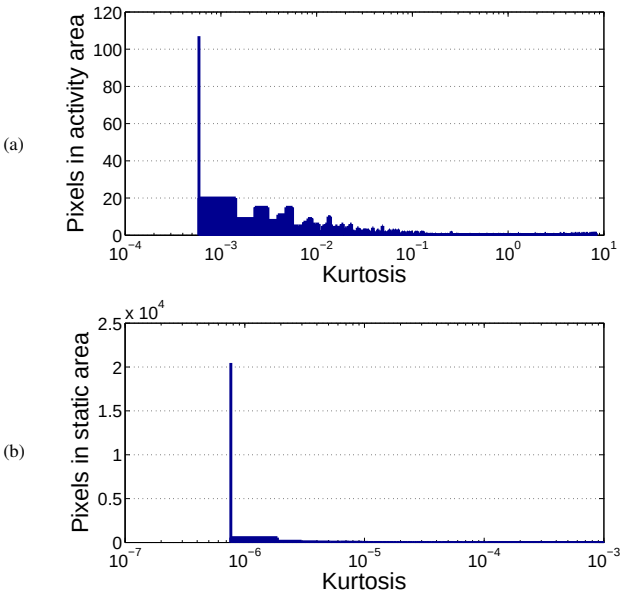


Fig. 2. Kurtosis estimates for the active (a) and static (b) pixels of the example of Fig. 1.

thresholding. Since there is no generally applicable way to determine the value of this threshold, the following well-performing value was selected after experimentation:

$$TH = \overline{|K_{ir}(x,y)|} + 4 \cdot \sigma_{|K_{ir}(x,y)|} , \quad (5)$$

where the arithmetic mean $\overline{|K_{ir}(x,y)|}$ and standard deviation $\sigma_{|K_{ir}(x,y)|}$ are calculated over all the kurtosis fields $K_{ir}(x,y)$ that have been computed for all shots s_i of a set of annotated video content that has been assembled for training purposes. Using this value, for every estimated kurtosis field a corresponding activity area mask is computed, according to the following equation:

$$A_{ir}(x,y) = \begin{cases} 1, & \text{if } |K_{ir}(x,y)| \geq TH \\ 0, & \text{otherwise} \end{cases} , \quad (6)$$

where $A_{ir}(x,y)$ is the estimated binary activity area mask.

In order to demonstrate how the kurtosis estimates provide reliable localization of active pixels, an indicative example showing the estimated kurtosis field and the corresponding binary activity area mask for a news domain video, and specifically a sequence showing an anchor presenting the news, is given in Fig. 1. In order to further examine the kurtosis values, the area of the active pixels for the same sequence is also manually determined. Using the manually obtained ground truth, the motion energy estimates of the 'active' pixels are separated from the corresponding estimates of the 'static' ones. For this particular video sequence, consisting of 288×352 pixel frames (total of 101376 pixels per frame), there are 13564 active and 87812 static pixels. In Fig. 2, two plots are illustrated; a histogram of the kurtosis of the manually determined active pixels' motion energy values, and a corresponding one of the static pixels' energy estimates. It is evident from this figure that the kurtosis of the active pixels obtains much higher values than that of the static

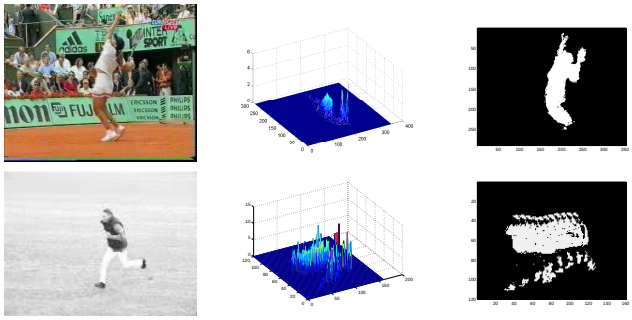


Fig. 3. Examples of kurtosis field and activity area mask computation for additional video sequences.

pixels. Specifically, its mean value over the entire sequence is 0.0281 for the active pixels, while for the static ones the respective value is 5.9391×10^{-7}. Hence, for this particular video sequence, the static pixels' mean kurtosis is equal to 0.0156% of the mean kurtosis of all frame pixels. Thus, it is shown from this example that the kurtosis fields can provide a reliable indicator for localizing the active pixels in a video sequence.

In Fig. 3, additional examples of kurtosis field estimation and activity area mask computation for video sequences from various domains are given. In particular, a segment of a broadcasted tennis game is depicted (first row), showing a player performing a servis hit. As can be seen, the only motion that is present in the sequence is the movement of the player performing the hit, which results in a kurtosis field with high values over the pixels that belong to the player's silhouette. Additionally, a scene of a person running is illustrated (second row), where the active pixels correspond to the displacement of the silhouette of the subject during the scene. In the above examples, the proposed approach has succeeded in ignoring noisy motion estimates originating mainly from random illumination changes.

C. Effect of Noise on Kurtosis-based Activity Area

In this section, the effect of noise on the computation of the activity area mask is examined. Specifically, the kurtosis of noisy motion energy estimates, i.e. when $M(x,y,t) = o(x,y,t) + n(x,y,t)$ (Eq. (2)), where $o(x,y,t)$ corresponds to the noiseless motion energy field and $n(x,y,t)$ to the additive noise, is calculated. It is assumed without loss of generality that $n(x,y,t)$ is zero-mean Gaussian and, for notational simplicity, the indices (x,y,t) in the equations that follow are not included; hence, $M = o + n$. The higher order moments of these variables are denoted by $m_{ok} = E[o^k]$, $m_{Mk} = E[M^k]$. Then:

$$(o+n)^4 = (o^2 + n^2 + 2o \cdot n)^2$$
$$= o^4 + n^4 + 6o^2n^2 + 4o^3n + 4o \cdot n^3, \quad (7)$$

and its expected value is:

$$E[(o+n)^4] = E[o^4 + n^4 + 6o^2n^2 + 4o^3n + 4o \cdot n^3]$$
$$= m_{o4} + m_{n4} + 6m_{o2}m_{n2} + 4m_{o3}m_{n1} + 4m_{o1}m_{n3}$$
$$= m_{o4} + m_{n4} + 6m_{o2}m_{n2}, \quad (8)$$

where it has been considered that $m_{n1} = 0$, and $m_{o1} = 0$, as the mean can be subtracted from the random variable o. Also:

$$E[(o+n)^2] = E[o^2 + n^2 + 2o \cdot n] = m_{o2} + m_{n2} \Rightarrow \quad (9)$$

$$E^2[(o+n)^2] = (m_{o2}+m_{n2})^2 = m_{o2}^2+m_{n2}^2+2m_{o2}m_{n2}. \quad (10)$$

The kurtosis is defined as $K_{o+n} = E[(o+n)^4]-3E^2[(o+n)^2]$, and Eqs. (8), (10) lead to:

$$K_{o+n} = m_{o4} + m_{n4} + 6m_{o2}m_{n2} -$$
$$-3m_{o2}^2 - 3m_{n2}^2 - 6m_{o2}m_{n2}$$
$$= m_{o4} + m_{n4} - 3m_{o2}^2 - 3m_{n2}^2 = K_o + K_n, \quad (11)$$

where the general definition of the kurtosis is used in the last equality. Eq. (11) has a central role in demonstrating the robustness of the kurtosis for the extraction of the activity area. Additive noise $n(x, y, t)$ is most often modeled as a Gaussian distribution. However the kurtosis of Gaussian random variables is equal to zero, so $K_n = 0$ and:

$$K_M \equiv K_{o+n} = K_o. \quad (12)$$

In other words, the kurtosis of the motion energy estimates remains unaffected by additive Gaussian noise. It should be noted that, even when the additive noise cannot be modeled by a Gaussian distribution, the kurtosis remains robust to noise and does not deviate significantly (if not at all) from its value in the noiseless case.

In order to demonstrate the robustness of the proposed approach, a comparison with a similar concept, namely that of Motion Energy Images (MEIs) [33] is presented. In [33], the pixels of activity are localized by direct thresholding of inter-frame differences and estimating the union of the resulting binary masks. Inter-frame illumination differences are reliable only for high-quality indoors videos, but can be replaced by flow estimates for noisier data (in this work the optical flow estimation algorithm cited in Section II-A was used). With respect to the MEIs calculation procedure, i.e. direct thresholding of the motion energy fields and consequently estimating the union of the resulting binary masks, it is obvious that they cannot eliminate the effect of additive Gaussian noise. Thus, the kurtosis-based activity areas are expected to be more robust and reliable in the presence of noise than the MEIs, as already shown theoretically.

In Fig. 4, indicative activity area and MEI estimation results, with noiseless data (rows 2 and 4) and in the presence of Gaussian additive noise (rows 3 and 5), are illustrated. Initially, a news broadcast video sequence showing an anchor presenting the news was used (first column). As can be seen, both the MEI and activity area provide an accurate localization of the pixels that move in the absence of noise. However, when additive noise is present, the MEI is seriously affected compared to the respective activity area, since it confuses noisy pixels with moving ones. Additionally, a comparison was conducted for a video of a person clapping (second column). As can be seen, the MEI contains significant amount of inaccurate estimations, i.e. it mistakes static pixels for moving ones, even in the absence of noise, whereas the corresponding activity area does not contain any artifacts. Moreover, when noise is added to the data, the activity area remains unaffected, while the MEI becomes significantly more erroneous. Additional videos that were also examined depict a person running (third column), a rally event from a tennis broadcast game (fourth column) and a player performing a servis hit (fifth column). As expected, the activity area remains more robust to the additive noise than the MEI, which looses significant amount of activity information. In order to numerically evaluate the robustness of the two methods, the MEI and the activity area were estimated for ten videos belonging to different domains (including those of Fig. 4), in the absence and in the presence of noise. The computed masks were compared with the manually extracted ground truth regions of active pixels, by evaluating the percentage of pixels that were correctly classified (either as active or as static). In the absence of noise, it was found that both algorithms exhibited high recognition rates (correct pixel classification rate $> 98\%$). On the other hand, when noise was present, the proposed kurtosis-based approach led to a 96.16% correct pixel classification rate, whereas the MEI one resulted in 88.51% of the pixels being correctly classified.

III. Motion Representation

The majority of the HMM-based analysis methods present in the relevant literature are focusing only at global- or camera-level motion representation approaches, as detailed in Section I. Nevertheless, local-level analysis of the motion signal can provide significant cues which, if suitably exploited, can facilitate in efficiently capturing the underlying semantics of the examined video. To this end, a new representation for providing local-level motion information to HMMs is presented here. It must be noted that the motion information processing described in this section applies to a single shot s_i at any time, thus indices i are omitted in this section for notational simplicity.

As already described in Section II, the kurtosis fields constitute a robust indicator for identifying pixels that undergo true motion. For representing the motion in the shot, it is reasonable to focus only on the pixels that are characterized as active in the corresponding activity area mask, i.e. the pixels where true motion is observed. These are more likely to bear significant information about the motion patterns that are discriminative for every supported class. In particular, for every computed activity area mask $A_r(x, y)$ a corresponding 'localized' mask $A_r^L(x_l, y_l)$, where $x_l \epsilon [x_r^{L0}, x_r^{L1}]$ $(1 \leq x_r^{L0} \leq x_r^{L1} \leq V_{dim})$ and $y_l \epsilon [y_r^{L0}, y_r^{L1}]$ $(1 \leq y_r^{L0} \leq y_r^{L1} \leq H_{dim})$, is estimated. This localized mask is defined as the axis-aligned minimum rectangle that includes all the active pixels of the respective $A_r(x, y)$, while maintaining the same aspect ratio. The corresponding 'localized' kurtosis field is denoted by $K_r^L(x_l, y_l)$, and comprises those pixels of $K_r(x, y)$ that belong to $A_r^L(x_l, y_l)$. The remainder of the motion analysis procedure considers only the $K_r^L(x_l, y_l)$ and $A_r^L(x_l, y_l)$.

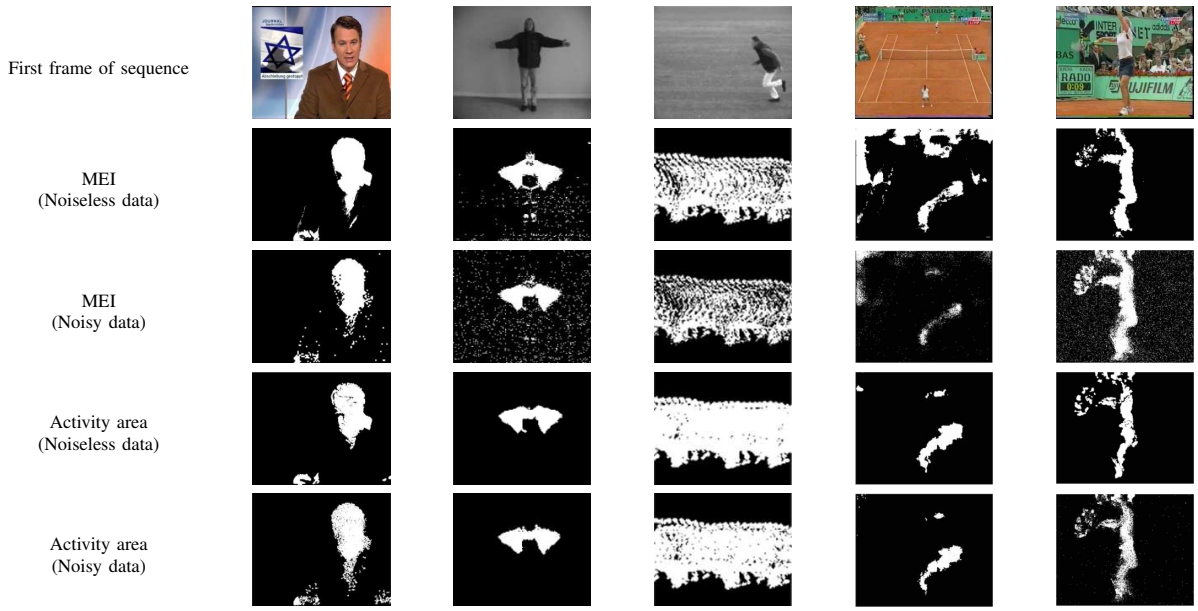


Fig. 4. MEI and activity area estimation with noiseless data (rows 2 and 4) and in the presence of Gaussian additive noise (rows 3 and 5).

A. Polynomial Approximation

The estimated localized kurtosis field, $K_r^L(x_l, y_l)$, is usually of high dimensionality, which decelerates the video processing, while motion information at this level of detail is not always required for the analysis purposes. Thus, it is down-sampled, according to the following equations:

$$\begin{aligned} K_r^{\Lambda}(x_\lambda, y_\lambda) &= K_r^L(x_d, y_d) \\ x_d = x_r^{L0} + \frac{2x_\lambda - 1}{2} \cdot V_s,\ \ y_d &= y_r^{L0} + \frac{2y_\lambda - 1}{2} \cdot H_s \\ x_\lambda = 1, ...D \ , \quad y_\lambda &= 1, ...D \\ V_s = \lfloor \frac{x_r^{L1} - x_r^{L0}}{D} \rfloor \ , \quad H_s &= \lfloor \frac{y_r^{L1} - y_r^{L0}}{D} \rfloor \end{aligned} \tag{13}$$

where $K_r^{\Lambda}(x_\lambda, y_\lambda)$ is the estimated down-sampled localized kurtosis field and H_s, V_s are the corresponding horizontal and vertical spatial sampling frequencies. As can be seen from Eq. (13), the dimensions of the down-sampled field are predetermined and set equal to D. It must be noted that if any of the sides of the localized kurtosis field is smaller than D (i.e. when $x_r^{L1} - x_r^{L0} < D$ or $y_r^{L1} - y_r^{L0} < D$), then $K_r^L(x_l, y_l)$ is interpolated so that its smaller side equals D, while maintaining the same aspect ratio as the original kurtosis field $K_r(x, y)$. Subsequently, the interpolated field, $\acute{K}_r^L(\acute{x}_l, \acute{y}_l)$, is down-sampled according to Eq. (13), where $\acute{K}_r^L(\acute{x}_d, \acute{y}_d)$ is used instead of $K_r^L(x_d, y_d)$. Interpolation is performed using the bilinear method.

According to the HMM theory [7], the set of sequential observation vectors that constitute an observation sequence need to be of fixed length and simultaneously of low-dimensionality. The latter constraint ensures the avoidance of HMM under-training occurrences. Thus, a compact and discriminative representation of motion features is required. For that purpose, the aforementioned $K_r^{\Lambda}(x_\lambda, y_\lambda)$ field, which actually represents a higher-order statistic of the motion energy distribution surface, is approximated by a 2D polynomial function, of the following form:

$$\begin{aligned} f(p, q) = \sum_{b,c} a_{bc} \cdot ((p - p_0)^b \cdot (q - q_0)^c) \ , \\ 0 \leq b, c \leq T \ \ and \ \ 0 \leq b + c \leq T \end{aligned} \tag{14}$$

where T is the order of the function, a_{bc} its coefficients and p_0, q_0 are defined as $p_0 = q_0 = \frac{D}{2}$. The approximation is performed using the least-squares method.

In Fig. 5, indicative examples of localized kurtosis field estimation and consequent approximation by a polynomial function are illustrated for various videos, showing the first frame of the sequence (first row), the estimated kurtosis field (second row), the resulting localized kurtosis field (third row) and its corresponding polynomial approximation $\widehat{K}_r^{\Lambda}(x_\lambda, y_\lambda)$ (row 4). As can be seen from this figure, the motion analysis localizes to the areas where increased motion activity is observed, while the subsequent polynomial approximation efficiently captures the most dominant local-level energy-distribution characteristics of the motion signal.

The proposed approximation of motion energy distribution, although quite simple, provides a very compact motion representation, since it estimates a low-dimensionality observation vector, while achieving to efficiently capture the most dominant motion characteristics of the examined video segment. Despite its sometimes rough approximation, the polynomial coefficients are experimentally shown to perform well in a number of different domains.

B. Spatial Attributes Extraction

While the estimated polynomial coefficients a_{bc} are used for approximating the computed localized kurtosis field $K_r^{\Lambda}(x_\lambda, y_\lambda)$, they do not capture spatial information regarding the size and position of the latter on the image grid. To this

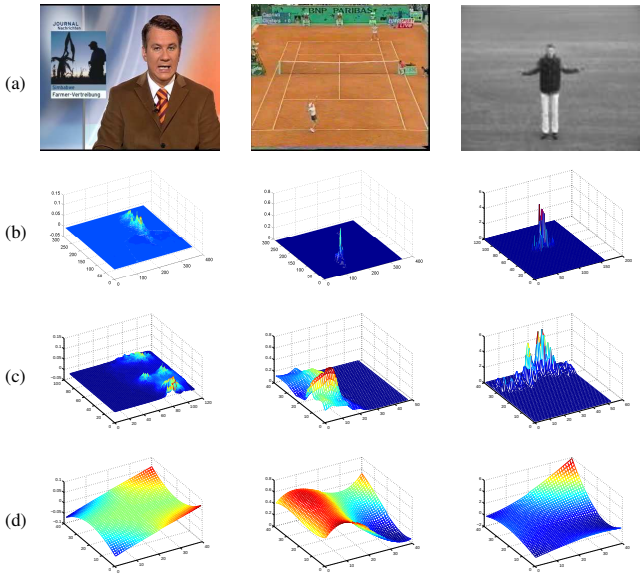


Fig. 5. Examples of localized kurtosis field estimation and approximation by polynomial function: (a) First frame of sequence, (b) Kurtosis field, (c) Localized kurtosis field and (d) Polynomial approximation.

end, three additional spatial features (relative area, center of gravity and displacement of center of gravity) are employed to compactly express this information. Moreover, a few other spatial features are also introduced to further emphasize particular spatial properties of the motion signal (like the orientation, rectangularity, etc. of the respective localized activity area), based on experimentation. All the aforementioned features, which constitute complementary information to the computed polynomial coefficients, are calculated from the estimated $A_r^L(x_l, y_l)$ mask. In particular, the employed features, which are extracted for every time interval w_r, are defined as follows:

- relative *area* of the estimated $A_r^L(x_l, y_l)$, which provides a quantitative measure of the size of the overall area where increased motion activity is observed:

$$area_r = \frac{(x_r^{L1} - x_r^{L0}) \cdot (y_r^{L1} - y_r^{L0})}{V_{dim} \cdot H_{dim}} \tag{15}$$

- *center of gravity* of the active pixels' region, which denotes the position of the activity area on the image grid:

$$\overrightarrow{CG_r} = (CG_r^0, CG_r^1)$$
$$CG_r^0 = \frac{\sum_{x_l} \sum_{y_l} x_l \cdot A_r^L(x_l, y_l)}{V_{dim} \cdot \sum_{x_l} \sum_{y_l} A_r^L(x_l, y_l)}$$
$$CG_r^1 = \frac{\sum_{x_l} \sum_{y_l} y_l \cdot A_r^L(x_l, y_l)}{H_{dim} \cdot \sum_{x_l} \sum_{y_l} A_r^L(x_l, y_l)} \tag{16}$$

- *displacement* of the active pixels' center of gravity in sequential time intervals:

$$\overrightarrow{DCG_r} = (CG_r^0 - CG_{r-1}^0, CG_r^1 - CG_{r-1}^1) \tag{17}$$

- *rectangularity*, which denotes how dense the active pixels' area is. It is defined as the percentage of the active pixels' *Minimum Bounding Rectangle* (MBR) that belongs to the respective $A_r^L(x_l, y_l)$:

$$rectangularity_r = \frac{\sum_{x_m} \sum_{y_m} A_r^L(x_m, y_m)}{(x_r^{M1} - x_r^{M0}) \cdot (y_r^{M1} - y_r^{M0})} \ , \tag{18}$$

where $x_m \epsilon [x_r^{M0}, x_r^{M1}]$, $y_m \epsilon [y_r^{M0}, y_r^{M1}]$, and $\{x_r^{M0}, x_r^{M1}, y_r^{M0}, y_r^{M1}\}$ denotes the MBR of the active pixels ($x_r^{L0} \leq x_r^{M0} \leq x_r^{M1} \leq x_r^{L1}$, $y_r^{L0} \leq y_r^{M0} \leq y_r^{M1} \leq y_r^{L1}$).

- *elongatedness* of the active pixels' MBR, which represents the thickness of the estimated activity area:

$$elongatedness_r = \frac{x_r^{M1} - x_r^{M0}}{y_r^{M1} - y_r^{M0}} \tag{19}$$

- *orientation*, which denotes the overall direction of the active pixels' region:

$$orientation_r = \frac{1}{2} \cdot \tan^{-1} \cdot \frac{2 \cdot \mu_{11}}{\mu_{20} - \mu_{02}} \ , \tag{20}$$

where $\mu_{11}, \mu_{20}, \mu_{02}$ are the corresponding *central moments* of $A_r^L(x_l, y_l)$.

- *accumulated active pixels ratio*, which is defined as the percentage of the total number of active pixels that are estimated from the beginning of shot s_i and are present in the current time interval w_r. This feature, which is particularly discriminative for periodic motions or movements with constant velocity, achieves to efficiently model the variation of motion intensity in time and is defined as follows:

$$R_r = \frac{E_r}{\sum_{\acute{r}=1}^{r} E_{\acute{r}}}, \ E_r = \sum_{x_l} \sum_{y_l} A_r^L(x_l, y_l) \tag{21}$$

The adopted spatial features express in a compact way particular attributes of the motion signal. Along with the introduced polynomial coefficients (Section III-A), they achieve to provide a more complete motion representation and efficiently capture the semantics that are present in the visual medium, facilitating in reaching improved classification performance.

IV. HMM-based Classification

As outlined in Section I, HMMs constitute a powerful statistical tool for solving problems that exhibit an inherent temporality, i.e. consist of a process that unfolds in time [7], [54]. The fundamental idea is that every process is made of a set of internal states and every state generates an observation when the process lies in that state. Thus, the sequential transition of the process among its constituent states generates a corresponding observation sequence. The latter is characteristic for every different process. It must be noted that a HMM requires a set of suitable training data for adjusting its internal structure, i.e. for efficiently modeling the process with which it is associated. At the evaluation stage, the HMM, which receives as input a possible observation sequence, estimates a posterior probability, which denotes the fitness of the input sequence to that model.

Under the proposed approach, HMMs are employed for associating every video shot with a particular semantic class, due to their increased applicability to modeling the temporal characteristics of the video sequence. In accordance to the

HMM theory, each class corresponds to a process that is to be modeled by an individual HMM and the features extracted from the video stream constitute the respective observation sequences. Specifically, since the polynomial coefficients and spatial attributes of the motion signal are estimated for a time interval w_{ir} of shot s_i (as detailed in Section III), they are used to form a single observation vector. These observation vectors for all w_{ir} of shot s_i form a respective shot observation sequence. Then, a set of J HMMs is employed, where an individual HMM is introduced for every defined class e_j, in order to perform the association of the examined shot, s_i, with the defined classes, e_j, based on the computed shot observation sequence. More specifically, each HMM receives the aforementioned observation sequence as input and estimates a posterior probability, which indicates the degree of confidence h_{ij} with which class e_j is associated with shot s_i. HMM implementation details are discussed in the experimental results section.

V. Experimental Results

In this section, experimental results from the application of the proposed approach in various domains, as well as comparative evaluation with other approaches in the literature, are presented. Although the approach is generic, i.e. it can be directly applied to any possible domain of concern without the need for domain-specific algorithmic modifications or adaptations, particular domains need to be selected for experimentation; to this end, the domains of tennis, news and volleyball broadcast video, and human action video are utilized in this work.

A. Tennis Domain

For experimentation in the domain of tennis broadcast video, four semantic classes of interest were defined, coinciding with four high-level semantic events that typically dominate a broadcasted game. These are:

- rally: when the actual game is played
- serve: is the event starting at the time that the player is hitting the ball to the ground, while he is preparing to serve, and finishes at the time the player performs the servis hit
- replay: when a particular incident of increased importance is broadcasted again, usually in slow motion
- break: when a break in the game occurs, i.e. the actual game is interrupted for example after a point is gained, and the camera may show the players resting or the audience

Then, a set of 8 videos showing professional tennis games from various international tournaments was collected. After the temporal segmentation algorithm of [50] was applied, a corresponding set of 886 shots was formed, which were manually annotated according to the respective event definitions. Out of the aforementioned videos, 4 (total of 499 shots; rally:191, serve:50, replay:31, break:227) were used for training the developed HMMs structure, while the remaining 4 (total of 387 shots; rally:130, serve:45, replay:20, break:192) were used for evaluation.

Every shot was further divided into a set of sequential time intervals of equal duration, as described in Section II-A. The duration of every interval, TW, was set to 0.40 sec, based on experimentation (the respective value for the news, volleyball and the action domains was set equal to 0.40 sec, 0.40 sec and 0.80 sec, respectively). It has been observed that small deviations from this value resulted into negligible changes in the overall detection performance. Then, for every resulting interval the corresponding kurtosis field, $K_{ir}(x, y)$, and activity area mask, $A_{ir}(x, y)$, were calculated, as detailed in Section II-B. Subsequently, the respective localized kurtosis field, $K^L_{ir}(x_l, y_l)$, and activity area mask, $A^L_{ir}(x_l, y_l)$, were computed with respect to the estimated active pixels. Local-level energy distribution-related information, as well as spatial attributes of the motion signal, were estimated from $K^L_{ir}(x_l, y_l)$ and $A^L_{ir}(x_l, y_l)$, as described in Sections III-A and III-B, respectively. A third order polynomial function was used for the approximation procedure, according to Eq. (14), since it produced the most accurate approximation results. The value of the parameter D in Eq. (13), which is used to define the horizontal, H_s, and vertical, V_s, spatial sampling frequencies, was set equal to 40. This value was shown to represent a good compromise between the need for time efficiency and accuracy of the polynomial approximation. Significantly lower values of D were shown to result into the generation of very few samples that could not be utilized for robust polynomial approximation. The motion features extracted for every time interval were used to form the motion observation sequence for the respective shot, which was in turn provided as input to the developed HMM structure in order to associate the shot with one of the supported classes, as described in Section IV. It must be noted that the values of every feature are normalized so that they have zero mean and standard deviation equal to one.

Regarding the HMM structure implementation details, fully connected first order HMMs, i.e. HMMs allowing all possible hidden state transitions, were utilized for performing the mapping of the low-level motion features to the high-level semantic classes. For every hidden state the observations were modeled as a mixture of Gaussians (a single Gaussian was used for every state). The employed Gaussian Mixture Models (GMMs) were set to have full covariance matrices for exploiting all possible correlations between the elements of each observation. Additionally, the Baum-Welch (or Forward-Backward) algorithm was used for training, while the Viterbi algorithm was utilized during the evaluation. Furthermore, the number of hidden states of the HMMs was considered as a free variable. The developed HMM structure was realized using the software libraries of [55].

In order to demonstrate and comparatively evaluate the efficiency of the proposed method, the following experiments were made:

1) application of the complete proposed approach (Sections II - IV)
2) application of the proposed approach without considering the spatial attributes of Section III-B
3) application of the proposed approach of Section II combined with a variant of the approach of III-A,

where $K_{ir}(x, y)$ is used in place of the localized kurtosis field $K_{ir}^{L}(x_l, y_l)$ (the spatial attributes presented in III-B are also not used)

4) application of the approach of Sections III and IV in combination with MHIs and MEIs [33] rather than the kurtosis fields and activity areas of Section II

5 - 7) application of the methods of [36], [39] and [40].

Experiments 1 and 2 are conducted in order to highlight the added value of incorporating spatial attributes of the motion signal in the classification process, along with local-level energy distribution-related information. Additionally, the performance of the proposed method is compared to the case when only global-level polynomial approximation of the kurtosis field is performed (experiment 3). In order to investigate the effectiveness of the proposed kurtosis field and activity area in capturing the characteristics of the motion signal, they are quantitatively evaluated against the *temporal template* motion representation approach presented in [33] (experiment 4). In particular, Bobick et al. introduce the Motion Energy Image (MEI) for denoting the pixels where motion has occurred in an image sequence. The latter is constructed by simple thresholding of the inter-frame difference at selected frames of the sequence and then computing the union of the resulting binary masks. Additionally, the Motion History Image (MHI) is proposed for describing the recency of motion and is produced by combining the aforementioned binary masks, where each mask is appropriately weighted with respect to its order in time. For realizing the performance comparison, an individual MEI and a corresponding MHI are computed for every estimated time interval w_{ir} (Section II-A). It must be noted that instead of utilizing inter frame-difference for computing the MHIs and MEIs, the estimated motion energy fields, $M(x, y, t)$, were employed. The latter are more robust to noise and provide a more accurate motion intensity field. Then, energy-distribution related information and spatial attributes of the motion signal are estimated from the computed MHI and MEI, respectively, as detailed in Sections III-A and III-B. Subsequently, class association is performed as described in Section IV. The proposed method is also comparatively evaluated against the representation approaches for providing motion information to HMM-based systems, with respect to semantic video analysis tasks, presented in [36], [39] and [40]. Specifically, Huang et al. consider the first four dominant motion vectors and their appearance frequencies, along with the mean and the standard deviation of motion vectors in the frame [36]. On the other hand, Gibert et al. estimate the principal motion direction of every frame [39], while Xie et al. calculate the motion intensity at frame level [40].

In Table I, quantitative class association results are given for the aforementioned experiments in the form of confusion matrices. The value of the overall classification accuracy is also given for each experiment. The latter is defined as the percentage of the video shots that are associated with the correct class. It has been regarded that $\arg\max_j(h_{ij})$ (Section IV) indicates the class e_j that is associated with shot s_i.

From the results presented in Table I, it can be seen that the proposed local-level representation approach for providing motion information to HMMs achieves an overall classification accuracy of 86.05%. More specifically, the class rally is recognized correctly at a high rate (98.46%), since it corresponds to a representative and distinguishable motion pattern. Additionally, classes serve and break also exhibit satisfactory results (82.22% and 81.77%, respectively). Regarding the recognition of replay, it presents a relatively low recognition rate (55.00%) and is mainly confused with class break. The latter is justified by the observation that replays are important incidents during the game that are broadcasted again usually in a close-up view and in slow-motion. Thus, they are expected to present similar local motion characteristics with class break. From the presented results, it can also be seen that the combination of local-level energy distribution-related information and spatial attributes of the motion signal leads to improved recognition results, compared to the case when only local-level energy distribution-related information is used. In particular, the incorporation of spatial features, extracted from the estimated localized activity area mask ($A_{ir}^{L}(x_l, y_l)$), leads to an increase of 10.86%, in the overall classification accuracy. Moreover, the detection of some classes (e.g. serve and replay) is particularly favored by the incorporation of the spatial features. The proposed motion representation approach is also advantageous compared to the case where only global-level energy distribution-related information is utilized. The latter is mainly due to the inefficacy of the global-level polynomial approximation to capture particular local characteristics of the motion signal. Additionally, it is shown that the proposed kurtosis field and activity area lead to increased detection rates for most of the supported classes, as well as in overall detection accuracy, compared to the corresponding MHI and MEI. Only for the class replay, the use of MHI and MEI leads to increased recognition performance. This is due to the fact that replay shots typically exhibit zoom-in and fading, and the combination of MHI and MEI is experimentally shown to be somewhat more suitable for modeling such effects. Furthermore, it should be noted that this large in percentage difference corresponds to only 2 more shots being misclassified, since the total number or replay shots in the real broadcast tennis video collection used for experimentation was relatively low and equal to 20. Finally, it can be observed that the proposed approach outperforms the algorithms of [36], [39] and [40], for all supported classes. This verifies that local-level analysis of the motion signal can lead to increased class association performance.

B. News Domain

For the domain of news broadcast video, the following semantic classes were defined:

- anchor: when the anchor person announces the news in a studio environment
- reporting: when live-reporting takes place or a speech/interview is broadcasted
- reportage: comprises of the displayed scenes, either indoors or outdoors, relevant to every broadcasted news item
- graphics: when any kind of graphics is depicted in the video sequence, including news start/end signals, maps, tables or text scenes

TABLE I
SEMANTIC CLASS ASSOCIATION RESULTS IN THE TENNIS DOMAIN FOR EXPERIMENTS (1) TO (7) (e_1: RALLY, e_2: SERVE, e_3: REPLAY AND e_4: BREAK)

Method	Actual Class	Associated Class			
		e_1	e_2	e_3	e_4
Proposed approach (1)	e_1	98.46%	0.77%	0.00%	0.77%
	e_2	2.22%	**82.22%**	0.00%	15.56%
	e_3	5.00%	0.00%	55.00%	40.00%
	e_4	5.73%	8.85%	3.65%	**81.77%**
	Overall Accuracy: **86.05%**				
Local-level polynomial approximation without spatial attributes (2)	e_1	97.69%	1.54%	0.00%	0.77%
	e_2	2.22%	55.56%	4.44%	37.78%
	e_3	0.00%	20.00%	35.00%	45.00%
	e_4	6.25%	16.67%	8.33%	68.75%
	Overall Accuracy: 75.19%				
Global-level polynomial approximation (3)	e_1	96.15%	3.08%	0.00%	0.77%
	e_2	4.44%	60.00%	8.89%	26.67%
	e_3	0.00%	10.00%	50.00%	40.00%
	e_4	5.21%	9.90%	11.46%	73.44%
	Overall Accuracy: 78.29%				
Proposed approach using method of [33] (4)	e_1	**99.23%**	0.77%	0.00%	0.00%
	e_2	4.44%	62.22%	22.22%	11.11%
	e_3	0.00%	10.00%	**65.00%**	25.00%
	e_4	13.54%	11.46%	33.33%	41.67%
	Overall Accuracy: 64.60%				
Method of [36] (5)	e_1	97.69%	0.77%	0.00%	1.54%
	e_2	6.67%	57.78%	8.89%	26.67%
	e_3	0.00%	15.00%	15.00%	70.00%
	e_4	8.33%	9.38%	4.17%	78.13%
	Overall Accuracy: 79.07%				
Method of [39] (6)	e_1	91.54%	8.46%	0.00%	0.00%
	e_2	35.56%	35.56%	11.11%	17.78%
	e_3	25.00%	20.00%	10.00%	45.00%
	e_4	18.23%	15.10%	10.42%	56.25%
	Overall Accuracy: 63.31%				
Method of [40] (7)	e_1	93.85%	6.15%	0.00%	0.00%
	e_2	6.67%	26.67%	51.11%	15.56%
	e_3	0.00%	20.00%	40.00%	40.00%
	e_4	8.33%	14.58%	36.46%	40.63%
	Overall Accuracy: 56.85%				

TABLE II
SEMANTIC CLASS ASSOCIATION RESULTS IN THE NEWS DOMAIN FOR EXPERIMENTS (1) TO (7) (e_1: ANCHOR, e_2: REPORTING, e_3: REPORTAGE AND e_4: GRAPHICS)

Method	Actual Class	Associated Class			
		e_1	e_2	e_3	e_4
Proposed approach (1)	e_1	95.45%	2.27%	2.27%	0.00%
	e_2	14.63%	63.41%	19.51%	2.44%
	e_3	4.44%	3.33%	**90.00%**	2.22%
	e_4	6.25%	0.00%	6.25%	**87.50%**
	Overall Accuracy: **86.83%**				
Local-level polynomial approximation without spatial attributes (2)	e_1	95.45%	4.55%	0.00%	0.00%
	e_2	39.02%	41.46%	19.51%	0.00%
	e_3	8.33%	10.56%	75.00%	6.11%
	e_4	12.50%	0.00%	12.50%	75.00%
	Overall Accuracy: 73.31%				
Global-level polynomial approximation (3)	e_1	90.91%	9.09%	0.00%	0.00%
	e_2	12.20%	**73.17%**	14.63%	0.00%
	e_3	5.00%	13.89%	80.00%	1.11%
	e_4	6.25%	0.00%	31.25%	62.50%
	Overall Accuracy: 79.72%				
Proposed approach using method of [33] (4)	e_1	**97.73%**	2.27%	0.00%	0.00%
	e_2	19.51%	63.41%	17.07%	0.00%
	e_3	10.00%	8.89%	77.22%	3.89%
	e_4	25.00%	0.00%	6.25%	68.75%
	Overall Accuracy: 77.94%				
Method of [36] (5)	e_1	86.44%	11.86%	0.00%	1.69%
	e_2	21.43%	57.14%	21.43%	0.00%
	e_3	5.75%	25.86%	66.67%	1.72%
	e_4	40.63%	3.13%	0.00%	56.25%
	Overall Accuracy: 68.60%				
Method of [39] (6)	e_1	18.18%	4.55%	0.00%	77.27%
	e_2	7.32%	17.07%	43.90%	31.71%
	e_3	1.67%	8.89%	80.00%	9.44%
	e_4	12.50%	6.25%	0.00%	81.25%
	Overall Accuracy: 61.21%				
Method of [40] (7)	e_1	52.27%	6.82%	0.00%	40.91%
	e_2	9.76%	39.02%	29.27%	21.95%
	e_3	6.11%	23.33%	63.89%	6.67%
	e_4	6.25%	18.75%	0.00%	75.00%
	Overall Accuracy: 59.07%				

Following a procedure similar to the one described in Section V-A, 24 videos of news broadcast from Deutsche Welle[1] were collected and the corresponding training and test sets were formed, comprising 338 (anchor:70, reporting:46, reportage:174, graphics:48) and 582 (anchor:91, reporting:85, reportage:374, graphics:32) shots, respectively.

In Table II, quantitative class association results are given for the news domain, where the same experiments and comparative evaluations as for the tennis domain were conducted. From this table, it can be seen that the proposed method accomplishes an overall classification accuracy of 86.83%. In particular, the classes anchor, reportage and graphics are correctly identified at high recognition rates (95.45%, 90.00% and 87.50%, respectively). Regarding the class reporting, although it exhibits satisfactory results (63.41%), it tends to be confused with anchor and reportage. The latter is caused by the fact that speech or interview occurrences may present similar motion patterns with anchor speaking or reportage scenes, respectively. Additionally, it can be observed that the proposed combination of local-level energy distribution-related information and spatial attributes of the motion signal is also advantageous for this particular domain, compared to the cases where either solely local-level energy distribution-related information is used or only global-level polynomial approximation of the kurtosis field is performed. In particular, the incorporation of the spatial features leads this time to an increase of 13.52%, in the overall classification accuracy, while the detection of classes reporting and reportage is particularly favored. On the other hand, only the classification rate of reporting is enhanced when global-level information is used, since the latter is proven to bear more discriminative information for this particular class. Moreover, the presented results show that the proposed kurtosis field and activity area lead to increased classification rates for most of the supported classes, as well as in overall classification accuracy, compared to the case when the MHI and MEI of [33] are utilized. Finally, it is shown that the proposed approach outperforms the methods of [36], [39] and [40] for all supported classes, similarly to the tennis domain.

[1] http://www.dw-world.de/

C. Volleyball Domain

For experimentation in the domain of volleyball broadcast video, four semantic classes of interest were defined, which coincide with four high-level semantic events that typically dominate a broadcasted game. In particular, the same events defined for the tennis domain, i.e. rally, serve, replay and break, were also used for this domain.

Following a procedure similar to the one described in Sections V-A and V-B, 14 videos of volleyball broadcast from the Beijing 2008 men's olympic volleyball tournament were collected and the corresponding training and test sets were formed, comprising 388 (rally:108, serve:55, replay:44, break:181) and 517 (rally:131, serve:80, replay:51, break:255) shots, respectively.

In Table III, quantitative class association results are given for the volleyball domain, where the same experiments and comparative evaluations as for the tennis and news domains were conducted. From the presented results, it can be seen that the proposed method achieves an overall classification accuracy of 88.39%. In particular, the classes rally, serve and break are correctly identified at high recognition rates (94.66%, 87.50% and 90.59%, respectively). Regarding the class replay, although it exhibits satisfactory results (62.75%), it is mainly confused with class break, similarly to the tennis domain results. Additionally, it can be observed that the proposed combination of local-level energy distribution-related information and spatial attributes of the motion signal results in improved recognition performance for most of the defined classes as well as overall, compared to the cases where either solely local-level energy distribution-related information is used or only global-level polynomial approximation of the kurtosis field is performed. Moreover, it leads to increased classification rates compared to the case when the MHI and MEI of [33] are utilized and also outperforms the methods of [36], [39] and [40], similarly to the tennis and news domains.

TABLE III
SEMANTIC CLASS ASSOCIATION RESULTS IN THE VOLLEYBALL DOMAIN FOR EXPERIMENTS (1) TO (7) (e_1: RALLY, e_2: SERVE, e_3: REPLAY AND e_4: BREAK)

Method	Actual Class	Associated Class e_1	e_2	e_3	e_4
Proposed approach (1)	e_1	**94.66%**	1.53%	0.00%	3.82%
	e_2	0.00%	**87.50%**	0.00%	12.50%
	e_3	0.00%	3.92%	62.75%	33.33%
	e_4	0.78%	7.45%	1.18%	**90.59%**
	Overall Accuracy: **88.39%**				
Local-level polynomial approximation without spatial attributes (2)	e_1	93.89%	0.76%	2.29%	3.05%
	e_2	0.00%	65.00%	1.25%	33.75%
	e_3	11.76%	3.92%	66.67%	17.65%
	e_4	3.92%	9.41%	7.06%	79.61%
	Overall Accuracy: 79.69%				
Global-level polynomial approximation (3)	e_1	93.89%	1.53%	0.00%	4.58%
	e_2	0.00%	**87.50%**	1.25%	11.25%
	e_3	1.96%	3.92%	**68.63%**	25.49%
	e_4	1.96%	5.49%	13.73%	78.82%
	Overall Accuracy: 82.98%				
Proposed approach using method of [33] (4)	e_1	xx.xx%	xx.xx%	xx.xx%	xx.xx%
	e_2	xx.xx%	xx.xx%	xx.xx%	xx.xx%
	e_3	xx.xx%	xx.xx%	**xx.xx%**	xx.xx%
	e_4	xx.xx%	xx.xx%	xx.xx%	**xx.xx%**
	Overall Accuracy: xx.xx%				
Method of [36] (5)	e_1	88.55%	8.40%	1.53%	1.53%
	e_2	3.75%	77.50%	6.25%	12.50%
	e_3	3.92%	15.69%	54.90%	25.49%
	e_4	0.78%	9.41%	19.61%	70.20%
	Overall Accuracy: 74.47%				
Method of [39] (6)	e_1	67.18%	6.87%	23.66%	2.29%
	e_2	5.00%	36.25%	30.00%	28.75%
	e_3	5.88%	43.14%	41.18%	9.80%
	e_4	32.55%	12.94%	19.61%	34.90%
	Overall Accuracy: 43.91%				
Method of [40] (7)	e_1	72.52%	19.08%	3.82%	4.58%
	e_2	16.25%	66.25%	2.50%	15.00%
	e_3	11.76%	23.53%	35.29%	29.41%
	e_4	6.67%	20.00%	42.75%	30.59%
	Overall Accuracy: 47.20%				

D. Human Action Domain

The performance of the proposed method was also evaluated for the task of human action recognition. It must be noted that for this particular task a series of dedicated approaches, exhibiting high recognition rates, have already been presented in the literature. The proposed method, although it does not exploit specific facts and characteristics that are only present in this domain (like human body silhouette extraction [56], body pose estimation [57], etc.), which can significantly facilitate the recognition procedure, nevertheless presents satisfactory results.

Regarding the set of semantic classes of interest, these coincide with the following human actions: *boxing*, *handclapping*, *handwaving*, *jogging*, *running* and *walking*. The video database of [58] was used for experimentation in this domain. In this database, each of the aforementioned actions was performed several times by 25 subjects in 4 different scenarios, namely outdoors, outdoors with scale variation, outdoors with different clothes and indoors. The corresponding training and test sets, which include instances of all scenarios and are defined in the database, comprise 760 (boxing:126, handclapping:124, handwaving:126, jogging:128, running:128, walking:128) and 864 (boxing:144, handclapping:144, handwaving:144, jogging:144, running:144, walking:144) shots, respectively.

For the experimental evaluation, the experiments 1-4 defined in Section V-A were conducted. Additionally, the performance of the proposed method was compared with the action recognition approach presented in [58], where Schüldt et al. utilize local space-time features for identifying human actions. Comparison with the approaches [36], [39] and [40] is omitted, since they perform worse than the dedicated method of [58].

In Table IV, quantitative class association results are shown. From the presented results, it can be seen that the proposed approach achieves an overall classification accuracy of 76.59%. Specifically, classes boxing, handclapping, handwaving and walking exhibit high recognition rates (93.01%, 78.47%, 85.42% and 90.28%, respectively). Regarding the classes running and jogging, they present relatively low recognition rates (58.89% and 53.61%, correspondingly) and they tend to be confused with classes jogging and walking, respectively. The latter is caused by the fact that individual subjects perform the same actions in different ways; a person may run as fast

as someone else is jogging, while the jogging action of an individual may be very similar to the walking one of an other person. Thus, it is reasonable, even for an individual human observer, that these two pairs of actions are confused. Moreover, it can be seen that generally actions that involve arm moves (boxing, handclapping, handwaving), are efficiently distinguished from more extensive body movements (jogging, running, walking). Additionally, the presented results indicate that the proposed local-level motion representation approach is again advantageous, compared to the case where only local-level energy distribution-related information or global-level polynomial approximation of the kurtosis field is utilized. In particular, the incorporation of spatial features leads to an increase of 15.29% in the overall class association accuracy, compared to the case where only local-level energy distribution-related information is used, while the detection of some classes, namely handwaving, jogging and running, is significantly favored. Moreover, it can also be seen that the proposed kurtosis field and resulting activity area prove to be more efficient in capturing the characteristics of the motion signal in this particular domain, compared to the motion representation approach of [33]. Finally, it is shown that the proposed method, although it has not been designed for the specific task of human action recognition, outperforms the method of [58] for most of the supported classes, as well as in overall classification accuracy. The latter demonstrates the robustness of the proposed method and its efficiency in achieving high recognition rates in domain-specific tasks, despite its generic nature.

E. Spatial Features Effectiveness

In order to further evaluate the contribution of the different kinds of spatial features presented in Section III-B, two additional experiments were conducted, with the spatial features divided to two sets: i) the ones defining the size and position of the localized activity area on the image grid (relative area, center of gravity and displacement of center of gravity), and ii) the remaining features that emphasize particular spatial attributes of the motion signal. For the tennis domain the combined use of the polynomial coefficients with the spatial features of set (i) leads to an increase of 7.50% in the overall classification performance, while with the features of set (ii) instead of those of set (i) the increase is 3.36%, compared to the performance reached by using solely the polynomial coefficients. The corresponding increase for the news, volleyball and human actions domains is 11.03%, 7.16%, 11.59% and 4.51%, 3.48%, 5.56%, respectively. Taking into account the classification results reported in Tables I-IV, it can be seen that both sets of features contribute to increased performance over the use of the polynomial coefficients alone, while the use of either one of the two sets leads to inferior performance compared to using both of them at the same time; the latter results to 10.86%, 13.52%, 8.70% and 15.29% increase of overall classification accuracy over the use of the polynomial coefficients alone, for the tennis, news, volleyball and human actions domain, respectively.

TABLE V
SEMANTIC CLASS ASSOCIATION RESULTS FOR DIFFERENT VALUES OF THE ORDER T OF THE POLYNOMIAL FUNCTION FOR EXPERIMENTS (1) AND (2) (OVERALL ACCURACY)

Method	T	Domain			
		Tennis	News	Volleyball	Human action
Proposed approach (1)	2	82.17%	86.12%	85.11%	73.46%
	3	**86.05%**	**86.83%**	**88.39%**	**76.59%**
	4	81.91%	85.05%	87.23%	67.09%
	5	82.17%	83.63%	87.04%	68.25%
	6	79.59%	83.27%	85.30%	65.12%
Local-level polynomial approximation without spatial attributes (2)	2	73.90%	70.11%	77.95%	55.04%
	3	**75.19%**	**73.31%**	**79.69%**	**61.30%**
	4	74.42%	72.24%	78.72%	58.29%
	5	74.94%	71.53%	78.34%	60.49%
	6	74.16%	70.46%	75.63%	54.00%

F. Effect of the Degree of the Polynomial Function

In order to investigate the effect of the introduced polynomial function's degree on the overall shot-class association performance, experiments 1 and 2 (defined in Section V-A) were conducted again for different values of the degree T (Eq. (14)) of the polynomial function. In particular, the shot-class association performance was evaluated when parameter T receives values ranging from 2 to 6. Values greater than 6 for parameter T resulted in significantly decreased recognition performance. The corresponding shot classification results for all supported domains are illustrated in Table V.

From the presented results it can be seen that the use of a 3^{rd} order polynomial function leads to the best overall performance for both experiments in all defined domains. Lower values of T $(T = 2)$ resulted in compact but at the same time very coarse kurtosis field approximation, which led to decreased shot classification accuracy. On the other hand, greater values of T $(T = 4, 5, 6)$, although resulted in more accurate approximation of the localized kurtosis fields compared to the case of $T = 3$, they led to the generation of observation vectors of significantly higher dimensionality. This fact, which generally hinders efficient HMM-based classification (as described in Section III-A), resulted again in decreased shot classification accuracy. It must be noted that for the cases of the 5^{th} and 6^{th} order polynomial function, HMM under-training occurrences were observed for both experiments in all domains, mainly due to the high dimensionality of the corresponding observation vectors. In order to perform HMM-based classification for these cases, Principal Component Analysis (PCA) was used for reducing the dimensionality of the observation vectors, as in [34], [59]. The target dimension of the PCA output was set equal to the dimension of the observation vector that is generated when using a 4^{th} order polynomial function, i.e. the highest value of T for which HMM under-training occurrences were not observed, while the resulting data were shown to still account for approximately 90% of the variance in the original data, which is typically the assumption in the relevant literature [34].

VI. CONCLUSIONS

In this paper, a generic approach to semantic video analysis that is based on the statistical processing and representation

TABLE IV
SEMANTIC CLASS ASSOCIATION RESULTS IN THE HUMAN ACTION DOMAIN FOR EXPERIMENTS (1) TO (4) AND COMPARISON WITH [58] (e_1: BOXING, e_2: HANDCLAPPING, e_3: HANDWAVING, e_4: JOGGING, e_5: RUNNING AND e_6: WALKING)

Method	Actual Class	Associated Class					
		e_1	e_2	e_3	e_4	e_5	e_6
Proposed approach (1)	e_1	93.01%	0.00%	2.80%	0.00%	0.00%	4.20%
	e_2	17.36%	78.47%	4.17%	0.00%	0.00%	0.00%
	e_3	3.47%	4.17%	**85.42%**	1.39%	1.39%	4.17%
	e_4	0.00%	0.00%	0.00%	53.61%	15.28%	31.11%
	e_5	0.00%	0.00%	0.00%	36.25%	58.89%	4.86%
	e_6	0.00%	0.00%	0.00%	9.72%	0.00%	**90.28%**
	Overall Accuracy: **76.59**%						
Local-level polynomial approximation without spatial attributes (2)	e_1	81.12%	9.79%	3.50%	0.00%	0.00%	5.59%
	e_2	6.94%	**80.56%**	11.81%	0.00%	0.00%	0.69%
	e_3	9.72%	17.36%	60.42%	8.33%	1.39%	2.78%
	e_4	1.39%	1.39%	0.69%	29.86%	6.25%	60.42%
	e_5	6.25%	2.08%	2.78%	38.19%	27.78%	22.92%
	e_6	1.39%	0.69%	0.00%	9.03%	0.69%	88.19%
	Overall Accuracy: 61.30%						
Global-level polynomial approximation (3)	e_1	87.41%	2.10%	3.50%	0.70%	1.40%	4.90%
	e_2	34.72%	64.58%	0.69%	0.00%	0.00%	0.00%
	e_3	14.58%	13.19%	64.58%	4.86%	2.78%	0.00%
	e_4	0.00%	0.00%	2.08%	25.69%	11.11%	61.11%
	e_5	0.00%	0.00%	0.00%	38.89%	31.25%	29.86%
	e_6	3.47%	2.08%	3.47%	2.78%	5.56%	82.64%
	Overall Accuracy: 59.33%						
Proposed approach using method of [33] (4)	e_1	81.12%	6.29%	5.59%	2.10%	0.00%	4.90%
	e_2	20.83%	56.94%	20.14%	0.00%	2.08%	0.00%
	e_3	11.81%	6.94%	81.25%	0.00%	0.00%	0.00%
	e_4	0.00%	0.00%	0.00%	**84.03%**	4.17%	11.81%
	e_5	0.00%	0.00%	0.00%	35.42%	**63.19%**	1.39%
	e_6	0.00%	0.00%	0.00%	15.97%	4.17%	79.86%
	Overall Accuracy: 74.39%						
Method of [58]	e_1	**97.92%**	0.69%	0.69%	0.00%	0.00%	0.69%
	e_2	35.42%	59.72%	3.47%	0.00%	0.00%	1.38%
	e_3	20.83%	4.86%	73.61%	0.00%	0.00%	0.69%
	e_4	0.00%	0.00%	0.00%	60.42%	16.67%	22.92%
	e_5	0.00%	0.00%	0.00%	38.89%	54.86%	6.25%
	e_6	0.00%	0.00%	0.00%	16.19%	0.00%	83.81%
	Overall Accuracy: 71.72%						

of the motion signal was presented. The proposed method employs the kurtosis of the optical flow motion estimates for identifying which motion values originate from true motion rather than measurement noise, resulting in the robust estimation of activity areas over a series of frames. Additionally, a new representation for providing local-level motion information to HMMs is presented. This is based on the elegant combination of energy distribution-related information with a complementary set of features that highlight particular spatial attributes of the motion signal. Experimental results in various domains demonstrated the efficiency of the proposed approach. Future work includes the examination of more sophisticated motion analysis techniques, as well as corresponding color and audio processing schemes, for realizing semantic video analysis based on multi-modal information.

ACKNOWLEDGMENT

The work presented in this paper was supported by the European Commission under contracts FP6-027685 MESH, FP6-045547 VIDI-Video, FP6-027538 BOEMIE and FP6-027026 K-Space.

REFERENCES

[1] S. Chang, "The holy grail of content-based media analysis," *Multimedia, IEEE*, vol. 9, no. 2, pp. 6–10, 2002.

[2] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, vol. 22, no. 12, pp. 1349–1380, 2000.

[3] A. Francois, R. Nevatia, J. Hobbs, R. Bolles, and J. Smith, "VERL: an ontology framework for representing and annotating video events," *Multimedia, IEEE*, vol. 12, no. 4, pp. 76–86, 2005.

[4] M. Naphade, J. Smith, J. Tesic, S. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis, "Large-scale concept ontology for multimedia," *Multimedia, IEEE*, vol. 13, no. 3, pp. 86–91, 2006.

[5] D. Sadlier and N. OConnor, "Event Detection in Field Sports Video Using Audio–Visual Features and a Support Vector Machine," *Circuits and Systems for Video Technology, IEEE Trans. on*, vol. 15, no. 10, p. 1225, 2005.

[6] W. Adams, G. Iyengar, C. Lin, M. Naphade, C. Neti, H. Nock, and J. Smith, "Semantic Indexing of Multimedia Content Using Visual, Audio, and Text Cues," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 2, pp. 170–185, 2003.

[7] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[8] C. Cheng and C. Hsu, "Fusion of audio and motion information on HMM-based highlight extraction for baseball games," *Multimedia, IEEE Trans. on*, vol. 8, no. 3, pp. 585–599, 2006.

[9] S. Liu, M. Xu, H. Yi, L. Chia, and D. Rajan, "Multimodal Semantic Analysis and Annotation for Basketball Video," *EURASIP Journal on Applied Signal Processing*, vol. 2006, 2006.

[10] C. Snoek and M. Worring, "Multimodal Video Indexing: A Review of the State-of-the-art," *Multimedia Tools and Applications*, vol. 25, no. 1, pp. 5–35, 2005.

[11] K. Peker, A. Alatan, and A. Akansu, "Low-level motion activity features for semantic characterization of video," in *Multimedia and Expo (ICME), IEEE Int. Conf. on*, 2000.

[12] X. Sun, B. Manjunath, and A. Divakaran, "Representation of motion activity in hierarchical levels for video indexing and filtering," in *Image Processing (ICIP), IEEE Int. Conf. on*, 2002.

[13] Y. Tan, D. Saur, S. Kulkami, and P. Ramadge, "Rapid estimation of camera motion from compressed video withapplication to video annotation," *Circuits and Systems for Video Technology, IEEE Trans. on*, vol. 10, no. 1, pp. 133–146, 2000.

[14] L. Duan, J. Jin, Q. Tian, and C. Xu, "Nonparametric motion characterization for robust classification of camera motion patterns," *Multimedia, IEEE Trans. on*, vol. 8, no. 2, pp. 323–340, 2006.

[15] R. Leonardi and P. Migliorati, "Semantic Indexing of Multimedia Documents," *Multimedia, IEEE*, pp. 44–51, 2002.

[16] Y. Ma and H. Zhang, "Motion Texture: A New Motion Based Video Representation," in *Pattern Recognition (ICPR), Proc. of Int. Conf. on*, vol. 16, 2002, pp. 548–551.

[17] B. Adams, C. Dorai, and S. Venkatesh, "Toward automatic extraction of expressive elements from motion pictures: tempo," *Multimedia, IEEE Trans. on*, vol. 4, no. 4, pp. 472–481, 2002.

[18] S. Dagtas, W. Al-Khatib, A. Ghafoor, R. Kashyap, P. Res, and B. Manor, "Models for motion-based video indexing and retrieval," *Image Processing, IEEE Trans. on*, vol. 9, no. 1, pp. 88–101, 2000.

[19] M. Roach, J. Mason, and M. Pawlewski, "Video genre classification using dynamics," in *Acoustics, Speech, and Signal Processing (ICASSP), IEEE Int. Conf. on*, vol. 3, 2001.

[20] C. Chen, J. Wang, J. Wang, and Y. Hu, "Motion Entropy Feature and Its Applications to Event-Based Segmentation of Sports Video," *EURASIP Journal on Advances in Signal Proc.*, vol. 2008.

[21] H. Pan, P. van Beek, and M. Sezan, "Detection of slow-motion replay segments in sports video forhighlights generation," in *Acoustics, Speech, and Signal Processing (ICASSP), IEEE Int. Conf. on*, vol. 3, 2001.

[22] L. Wang, X. Liu, S. Lin, G. Xu, and H. Shum, "Generic slow-motion replay detection in sports video," in *Image Processing (ICIP), IEEE Int. Conf. on*, vol. 3, 2004.

[23] N. Rea, R. Dahyot, and A. Kokaram, "Modeling high level structure in sports with motion driven HMMs," in *Acoustics, Speech, and Signal Processing (ICASSP), IEEE Int. Conf. on*, vol. 3, 2004.

[24] H. Chen, D. Chen, and S. Lee, "Object based video similarity retrieval and its application to detecting anchorperson shots in news video," in *Multimedia Software Engineering, Int. Symposium on*, 2003, pp. 172–179.

[25] L. Chaisorn, T. Chua, and C. Lee, "The segmentation of news video into story units," in *Multimedia and Expo (ICME), IEEE Int. Conf. on*, vol. 1, 2002.

[26] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *Circuits and Systems for Video Technology, IEEE Trans. on*, vol. 18, no. 11, pp. 1473–1488, Nov. 2008.

[27] W. Hu, T. Tan, L. Wang, and S. Maybank, "A Survey on Visual Surveillance of Object Motion and Behaviors," *Systems, Man, and Cybernetics-Part C, IEEE Trans. on*, vol. 34, no. 3, 2004.

[28] N. Cuntoor, B. Yegnanarayana, and R. Chellappa, "Activity Modeling Using Event Probability Sequences," *Image Processing, IEEE Trans. on*, vol. 17, no. 4, pp. 594–607, 2008.

[29] A. Chan and N. Vasconcelos, "Modeling, Clustering, and Segmenting Video with Mixtures of Dynamic Textures," *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, pp. 909–926, 2008.

[30] R. Fablet and P. Bouthemy, "Motion Recognition Using Nonparametric Image Motion Models Estimated from Temporal and Multiscale Co-Occurrence Statistics," *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, pp. 1619–1624, 2003.

[31] W. Hu, X. Xiao, Z. Fu, D. Xie, T. Tan, and S. Maybank, "A System for Learning Statistical Motion Patterns," *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, pp. 1450–1464, 2006.

[32] Y. Rui and P. Anandan, "Segmenting visual actions based on spatio-temporal motion patterns," *Computer Vision and Pattern Recognition (CVPR), IEEE Int. Conf. on*, vol. 1, pp. 111–118, 2000.

[33] A. Bobick and J. Davis, "The Recognition of Human Movement Using Temporal Templates," *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, vol. 23, pp. 257–267, 2001.

[34] M. Barnard and J. Odobez, "Sports Event Recognition Using Layered HMMS," *Multimedia and Expo (ICME), IEEE Int. Conf. on*, pp. 1150–1153, 2005.

[35] G. Xu, Y. Ma, H. Zhang, and S. Yang, "An HMM-based framework for video semantic analysis," *Circuits and Systems for Video Technology, IEEE Trans. on*, vol. 15, no. 11, pp. 1422–1433, 2005.

[36] J. Huang, Z. Liu, and Y. Wang, "Joint scene classification and segmentation based on hidden Markov model," *Multimedia, IEEE Trans. on*, vol. 7, no. 3, pp. 538–550, 2005.

[37] J. Wang, C. Xu, and E. Chng, "Automatic Sports Video Genre Classification using Pseudo-2D-HMM," *Proc. of the 18th Int. Conf. on Pattern Recognition (ICPR)*, pp. 778–781, 2006.

[38] Y. Haoran, D. Rajan, and C. Liang-Tien, "An efficient video classification system based on HMM in compressed domain," *Proc. of the Joint Conf. of the Int. Conf. on Information, Communications and Signal Processing, and the Pacific Rim Conf. on Multimedia*, vol. 3, 2003.

[39] X. Gibert, H. Li, and D. Doermann, "Sports video classification using HMMS," *Multimedia and Expo (ICME), IEEE Int. Conf. on*, 2003.

[40] L. Xie, P. Xu, S. Chang, A. Divakaran, and H. Sun, "Structure analysis of soccer video with domain knowledge and hidden Markov models," *Pattern Recog. Letters*, vol. 25, no. 7, pp. 767–775, 2004.

[41] M. Welling, "Robust Higher Order Statistics," in *Proc. Int. Workshop Artif. Intell. Statist.(AISTATS*, 2005, pp. 405–412.

[42] C. Chi, C. Chen, C. Chen, and C. Feng, "Batch processing algorithms for blind equalization using higher-order statistics," *Signal Processing Magazine, IEEE*, vol. 20, no. 1, pp. 25–49, 2003.

[43] J. Kim, J. Choi, J. Yi, and M. Turk, "Effective Representation Using ICA for Face Recognition Robust to Local Distortion and Partial Occlusion," *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, pp. 1977–1981, 2005.

[44] C. Papadias, "Globally convergent blind source separation based on a multiuser kurtosis maximization criterion," *Signal Processing, IEEE Trans. on*, vol. 48, no. 12, pp. 3508–3519, 2000.

[45] L. Zelnik-Manor and M. Irani, "Event-based analysis of video," *Computer Vision and Pattern Recognition (CVPR), IEEE Int. Conf. on*, vol. 2, pp. 123–130, 2001.

[46] E. Lee and D. Messerschmitt, *Digital Communication*. Kluwer Academic Publishers, 1994.

[47] E. Moulines and J. Cardoso, "Second-order versus fourth-order music algorithms: an asymptotical statistical analysis," in *Proc. IEEE Signal Processing Workshop on Higher-Order Statistics, Chamrousse, France, June*, 1991.

[48] B. Sinha, "Detection of Multivariate Outliers in Elliptically Symmetric Distributions," *The Annals of Statistics*, vol. 12, no. 4, pp. 1558–1565, 1984.

[49] G. T. Papadopoulos, V. Mezaris, I. Kompatsiaris, and M. G. Strintzis, "Accumulated Motion Energy Fields Estimation and Representation for Semantic Event Detection," *Proc. Int. Conf. on Image and Video Retrieval (CIVR)*, 2008.

[50] V. Kobla, D. Doermann, and K. Lin, "Archiving, indexing, and retrieval of video in the compressed domain," *Multimedia Storage and Archiving Systems, SPIE Conf. on*, vol. 2916, pp. 78–89, 1996.

[51] M. Proesmans, L. Van Gool, E. Pauwels, and A. Oosterlinck, "Determination of Optical Flow and its Discontinuities using Non-Linear Diffusion," *European Conf. on Computer Vision (ECCV)*, 1994.

[52] R. Gonzalez and R. Woods, *Digital Image Processing*. Prentice Hall, 2007.

[53] G. Giannakis and M. Tsatsanis, "Time-domain tests for Gaussianity and time-reversibility," *Signal Processing, IEEE Trans. on*, vol. 42, no. 12, pp. 3460–3472, 1994.

[54] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. Wiley-Interscience, 2000.

[55] "http://htk.eng.cam.ac.uk/," *Hidden Markov Model Toolkit, HTK*.

[56] S. Zhao and H. Lee, "Human Silhouette Extraction based on HMM," *Pattern Recognition (ICPR), Int. Conf. on*, pp. 994–997, 2006.

[57] A. Mittal, L. Zhao, and L. Davis, "Human body pose estimation using silhouette shape analysis," *Advanced Video and Signal Based Surveillance, IEEE Int. Conf. on*, pp. 263–270, 2003.

[58] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach," *Pattern Recognition (ICPR), Int. Conf. on*, vol. 3, 2004.

[59] S. Fu, R. Gutierrez-Osuna, A. Esposito, P. Kakumanu, and O. Garcia, "Audio/visual mapping with cross-modal hidden Markov models," *Multimedia, IEEE Trans. on*, vol. 7, no. 2, pp. 243–252, 2005.