Semantic Video Analysis Based on Estimation and Representation of Higher-Order Motion Statistics*

G. Th. Papadopoulos[1,2], A. Briassouli[2], V. Mezaris[2], I. Kompatsiaris[2] and M. G. Strintzis[1,2]
[1]Information Processing Lab., Electrical & Comp. Eng. Dep., Aristotle Univ. of Thessaloniki, Greece
[2]Informatics and Telematics Institute, Centre for Research and Technology Hellas, Greece
mail: {papad, abria, bmezaris, ikom}@iti.gr , strintzi@eng.auth.gr

Abstract

In this paper, a generic motion-based approach to semantic video analysis is presented. The examined video is initially segmented into shots and for every resulting shot appropriate motion features are extracted at fixed time intervals. Then, Hidden Markov Models (HMMs) are employed for performing the association of each shot with one of the semantic classes that are of interest in any given domain. Regarding the motion feature extraction procedure, higher order statistics of the motion estimates are calculated and a new representation for providing local-level motion information to HMMs is presented. The latter is based on the combination of energy distribution-related information and spatial attributes of the motion signal. Experimental results as well as comparative evaluation from the application of the proposed approach in the domain of news broadcast video are presented.

1. Introduction

Given the continuously increasing amount of video content generated everyday and the richness of the available means for sharing and distributing it, the need for efficient and advanced methodologies regarding video manipulation emerges as a challenging and imperative issue. To this end, several approaches have been proposed in the literature regarding the tasks of indexing, searching, retrieval, as well as personalized delivery of video content [1].

More recently, the fundamental principle of shifting video manipulation techniques towards the processing of the visual content at a semantic level has been widely adopted, thus attempting to bridge the so called *semantic gap* [11]. Among the video analysis methodologies of the latter category, approaches that exploit *a priori* knowledge have been particularly favored and have so far exhibited promising results.

Knowledge-assisted video analysis techniques have been dominated by the usage of Machine Learning (ML) algorithms. ML-based approaches utilize probabilistic methods for acquiring the appropriate implicit knowledge that will enable the mapping of the low-level audio-visual data to high-level semantic concepts and entities. In [4], a HMM-based system is proposed for performing joint scene classification and video temporal segmentation. Additionally, in [9], Support Vector Machines (SVMs) are employed for detecting semantically meaningful events in broadcast video of multiple field sports. Although many methods have already been presented for realizing knowledge-assisted video analysis, most of them are only limited to domain specific applications, i.e. they exploit specific facts and characteristics that are only present in a single domain, thus failing to effectively handle the problem of semantic video analysis at a more generic level.

In this paper, a generic motion-based approach to semantic video analysis, making use of ML algorithms for implicit knowledge acquisition, is presented. The examined video is initially segmented into shots and for every resulting shot appropriate motion features are extracted at fixed time intervals, thus forming a *motion observation sequence*. Then, HMMs are employed for performing the association of each shot with one of the supported semantic classes based on its formed observation sequence. Regarding the motion feature extraction procedure, higher order statistics of the motion estimates are calculated and result into a *kurtosis field*. The latter is highly sensitive to outliers, and hence provides a robust indication of which motion values originate from true motion rather than measurement noise. Additionally, a new representation for providing local-level motion information to HMMs is presented. This representation is based on the combination of energy distribution-related information and spatial attributes of the motion signal, for efficiently capturing the semantics present in the visual medium.

*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.

978-0-7695-3444-2/08 $25.00 © 2008 IEEE
DOI 10.1109/SMAP.2008.22

The paper is organized as follows: The video pre-processing steps are described in Section 2. Section 3 presents the statistical analysis of the motion signal. Section 4 details the extraction of the motion features. Section 5 discusses how HMMs are utilized for performing motion-based classification. Experimental results and comparative evaluation from the application of the proposed approach in the news broadcast domain are presented in Section 6, and conclusions are drawn in Section 7.

2. Video Pre-Processing

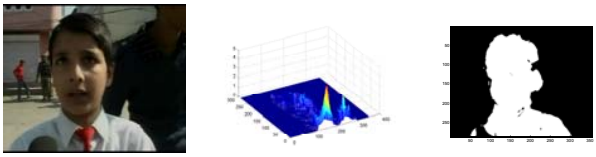
The examined video sequence is initially segmented into a set of shots, denoted by $S = \{s_i, i = 1, ...I\}$, which constitute the elementary image sequences of video; under the proposed approach each shot will be associated with one of the supported semantic classes, denoted by $E = \{e_j,\ j = 1, ..J\}$, on the basis of its semantic contents. For shot detection, the algorithm of [5] is used, mainly due to its low computational complexity. After shot segmentation, each shot s_i is further divided into a set of sequential 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, making use of the optical flow estimation algorithm of [7]. From the computed motion field a corresponding motion energy field is calculated, according to the following equation:

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

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 justified by the observation that often the latter provides more appropriate information for motion-based recognition problems [6].

3. Statistical Motion Analysis

The motion energy estimates at each pixel represent changes in illumination that originate either from measurement noise, or from pixel displacement (true motion) and



First frame of sequence | Kurtosis field | Activity area mask

Figure 1. Example of kurtosis field and activity area mask computation

measurement noise. This can be expressed as the following hypotheses:

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

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. 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 [2]. The classical measure of a random variable's non-gaussianity is its kurtosis, defined by:

$$kurt(\psi) = E[\psi^4] - 3(E[\psi^2])^2, \quad (3)$$

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

Although the Gaussian model is only an approximation of the unknown noise in the motion estimates, the kurtosis remains appropriate for detecting the true velocity measurements. This is because they appear as outliers, and in [10] 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.

In order to determine the pixels that undergo true motion within a particular time interval w_{ir}, the kurtosis value of every pixel is calculated, taking into account the computed motion energy estimates, $M(x, y, t)$, over all frames that belong to the specific time interval, according to the following equation:

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

where $K_{ir}(x, y)$ is the estimated kurtosis field and the expectations $E[\cdot]$ are approximated by the corresponding arithmetic means.

From the kurtosis field computation procedure, it is evident that pixels which undergo true motion present significantly higher kurtosis values, compared to the pixels that exhibit only measurement noise. Thus, the kurtosis fields can be considered as a reliable indicator of pixels' activity, allowing the distinction between 'active' and 'static' pixels by simple 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 sequence is given in Fig. 1. It is evident from this figure that the kurtosis of the active pixels obtains much higher values than that of the static pixels.

4. Motion Features Extraction

The majority of the HMM-based analysis methods present in the relevant literature are focusing only at global- or camera-level motion representation approaches [3][12]. 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 3, the kurtosis fields constitute a robust indicator for identifying pixels that undergo true motion. Hence, 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, since 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. The latter is defined as the minimum rectangle that includes all the active pixels of the respective $A_r(x,y)$, and maintains the same aspect ratio and orientation as the original $A_r(x,y)$. 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 $A_r^L(x_l,y_l)$.

4.1. 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 consequently down-sampled, according to the following equations:

$$\begin{aligned} K_r^\Lambda(x_\lambda,y_\lambda) &= K_r^L(x_d,y_d) \\ x_d &= x_r^{L0} + \frac{2x_\lambda - 1}{2} \cdot V_{step} \\ y_d &= y_r^{L0} + \frac{2y_\lambda - 1}{2} \cdot H_{step} \\ x_\lambda &= 1,...D \ , \quad y_\lambda = 1,...D \\ V_{step} = \lfloor \frac{x_r^{L1} - x_r^{L0}}{D} \rfloor , & \quad H_{step} = \lfloor \frac{y_r^{L1} - y_r^{L0}}{D} \rfloor \end{aligned} \quad (7)$$

where $K_r^\Lambda(x_\lambda,y_\lambda)$ is the estimated down-sampled localized kurtosis field and H_{step}, V_{step} are the corresponding horizontal and vertical spatial sampling frequencies. As can be seen from Eq. (7), the dimensions of the down-sampled field are predetermined and set equal to D. It must be noted that if $x_r^{L1} - x_r^{L0} < D$ or $y_r^{L1} - y_r^{L0} < D$, $K_r^L(x_l,y_l)$ is interpolated, before being down-sampled, following the bilinear method, so that the condition $x_r^{L1} - x_r^{L0}, y_r^{L1} - y_r^{L0} \geq D$ is satisfied.

According to the HMM theory [8], 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:

$$f(p,q) = \sum_{b,c} a_{bc} \cdot ((p-p_0)^b \cdot (q-q_0)^c) ,$$

$$0 \leq b, c \leq T \ \ and \ \ 0 \leq b + c \leq T \tag{8}$$

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. 2, an indicative example of localized kurtosis field estimation and approximation by a polynomial function is illustrated for a news video sequence. As can be seen from this figure, the polynomial approximation efficiently captures the most dominant motion characteristics.

4.2. Spatial Attributes Extraction

The estimated polynomial coefficients a_{bc} do not encompass information about the spatial properties of the motion signal (e.g. size and position of the computed $K_r^{\Lambda}(x_\lambda, y_\lambda)$). Hence, in this section, an additional set of features is defined for capturing the spatial attributes of the latter. These features, which constitute complementary information to the computed polynomial coefficients, highlight particular spatial attributes of the motion signal and 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 is calculated as follows:

$$area_r = \frac{(x_r^{L1} - x_r^{L0}) \cdot (y_r^{L1} - y_r^{L0})}{V_{dim} \cdot H_{dim}} \tag{9}$$

- *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 $A_r^L(x_l, y_l)$, and is estimated according to the following equation:

$$rectangularity_r = \frac{\sum_{x_m} \sum_{y_m} A_r^L(x_m, y_m)}{(x_r^{M1} - x_r^{M0}) \cdot (y_r^{M1} - y_r^{M0})} \ , \tag{10}$$

  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:

$$elongatedness_r = \frac{x_r^{M1} - x_r^{M0}}{y_r^{M1} - y_r^{M0}} \tag{11}$$

- *orientation*, which denotes the overall direction of the active pixels' region and is estimated as follows:

$$orientation_r = \frac{1}{2} \cdot \tan^{-1} \cdot \frac{2 \cdot \mu_{11}}{\mu_{20} - \mu_{02}} \ , \tag{12}$$

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

- *center of gravity* of the active pixels' region, which is calculated according to the following equations:

$$\begin{aligned} \overrightarrow{CG_r} &= (CG_r^0, CG_r^1) \\ CG_r^0 &= \frac{\sum_{x_l} \sum_{y_l} x_l \cdot A_r^L(x_l, y_l)}{V_{dim} \cdot \sum_{x_l} \sum_{y_l} A_r^L(x_l, y_l)} \\ CG_r^1 &= \frac{\sum_{x_l} \sum_{y_l} y_l \cdot A_r^L(x_l, y_l)}{H_{dim} \cdot \sum_{x_l} \sum_{y_l} A_r^L(x_l, y_l)} \end{aligned} \tag{13}$$

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

$$\overrightarrow{DCG_r} = (CG_r^0 - CG_{r-1}^0, CG_r^1 - CG_{r-1}^1) \tag{14}$$

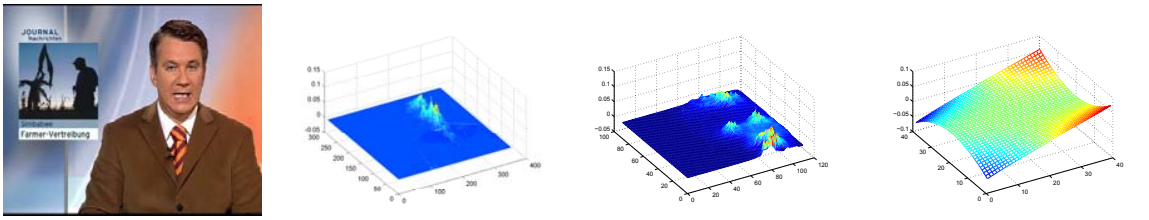
- *accumulated active pixels ratio*, which is defined as the percentage of the total number of active pixels that are estimated from the beginning of shot s_i and are present in the current time interval w_r. This feature achieves to efficiently model the variation of motion intensity in time and is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

$$\begin{aligned} R_r &= \frac{E_r}{\sum_{r=1}^{r} E_r} \\ E_r &= \sum_{x_l} \sum_{y_l} A_r^L(x_l, y_l) \end{aligned} \tag{15}$$

5. HMM-based Classification

HMMs constitute a powerful statistical tool for solving problems that exhibit an inherent temporality, i.e. consist of a process that unfolds in time [8]. 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 characteristic observation sequence. It must be noted that a HMM requires a set of suitable training data for adjusting its internal structure. At the evaluation stage, a 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. 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 4), 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.



First frame of sequence | Kurtosis field | Localized kurtosis field | Polynomial approximation

Figure 2. Example of localized kurtosis field approximation with polynomial function

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.

6. Experimental Results

In this section experimental results from the application of the proposed method, as well as comparative evaluation with other approaches in the literature, are presented. Although the method is generic, a domain needs to be selected for experimentation; to this end, the domain of news broadcast video is utilized in this work. For the selected domain, 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) and *graphics* (when any kind of graphics is depicted in the video sequence, including news start/end signals, maps, tables or text scenes).

Then, a set of 24 videos of news broadcast from Deutsche Welle[1] was collected. After the temporal segmentation algorithm of [5] was applied, a corresponding set of 924 shots was formed, which were manually annotated according to the class definitions already described. From the aforementioned videos 8 of them (342 shots) were used for training the developed HMM structure and the remaining 16 (582 shots) were used for evaluation.

Every shot was further divided into a set of sequential time intervals of equal duration, as described in Section 2. The duration of every interval, TW, was set to 0.40sec based on experimentation. It has been observed that small deviations from this value ($\pm 20\%$) resulted into negligible changes in the overall detection performance. Then, for every resulting interval the respective kurtosis field and activity area mask were calculated, as outlined in Section 3. Subsequently, local-level energy distribution-related information, as well as spatial attributes of the motion signal, were estimated, as detailed in Section 4. A third order polynomial function was used for the approximation procedure (Eq. (8)), since it produced the most accurate approximation results compared to the cases where e.g. a second or a fourth order polynomial function was used. The value of parameter D in Eq. (7) was set equal to 40. This value was shown to represent a good compromise between the need for time efficiency and effective 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 5.

Regarding the HMM structure implementation details, fully connected first order HMMs were utilized. For every hidden state the observations were modeled as a mixture of Gaussians, which were set to have full covariance matrices. Additionally, the Baum-Welch (or Forward-Backward) algorithm was used for training, while the Viterbi algorithm was utilized during the evaluation. The number of hidden states of the HMMs was considered as a free variable.

In Table 1, quantitative class association results are given in the form of the calculated confusion matrices from the application of the proposed approach in videos of the selected domain, when: a) only energy distribution-related information is utilized (first row), b) spatial attributes of the motion signal are also used (second row). Additionally, the value of the overall classification accuracy is also given, which is defined as the percentage of the video shots that are correctly classified. It has been regarded that $\arg\max_j(h_{ij})$ indicates the class e_j that is associated with shot s_i.

From the results presented in Table 1, it can be seen that the combination of energy distribution-related information and spatial attributes of the motion signal leads to better recognition results, compared to the case when only distribution-related information is used. Additionally, it is observed that the proposed motion feature extrac-

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

tion approach for providing motion information to HMMs achieves an overall classification accuracy of 86.83%, while most of the supported classes are correctly identified at high recognition rates. With respect to 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.

The performance of the proposed method is also compared with the motion representation approaches for providing motion information to HMM-based systems presented in [4], [3] and [12]. Specifically, Huang et al. considers the first four dominant motion vectors and their appearance frequencies, along with the mean and the standard deviation of motion vectors in the frame [4]. On the other hand, Gibert et al. estimates the principal motion direction of every frame [3], while Xie et al. calculates the motion intensity at frame level [12]. From the presented results, it can be easily observed that the proposed approach outperforms the aforementioned algorithms for all supported classes as well as in overall classification accuracy. This verifies that local-level analysis of the motion signal can lead to increased class association performance.

Table 1. Semantic class association results

Method	Associated Class				Actual Class
	Anchor	Reporting	Reportage	Graphics	
Proposed approach (using only a_{bc})	**95.45**%	4.55%	0.00%	0.00%	Anchor
	39.02%	41.46%	19.51%	0.00%	Reporting
	8.33%	10.56%	75.00%	6.11%	Reportage
	12.50%	0.00%	12.50%	75.00%	Graphics
	Overall Accuracy				73.31%
Proposed approach	**95.45**%	2.27%	2.27%	0.00%	Anchor
	14.63%	**63.41**%	19.51%	2.44%	Reporting
	4.44%	3.33%	**90.00**%	2.22%	Reportage
	6.25%	0.00%	6.25%	**87.50**%	Graphics
	Overall Accuracy				**86.83**%
Method of [4]	86.44%	11.86%	0.00%	1.69%	Anchor
	21.43%	57.14%	21.43%	0.00%	Reporting
	5.75%	25.86%	66.67%	1.72%	Reportage
	40.63%	3.13%	0.00%	56.25%	Graphics
	Overall Accuracy				68.60%
Method of [3]	18.18%	4.55%	0.00%	77.27%	Anchor
	7.32%	17.07%	43.90%	31.71%	Reporting
	1.67%	8.89%	80.00%	9.44%	Reportage
	12.50%	6.25%	0.00%	81.25%	Graphics
	Overall Accuracy				61.21%
Method of [12]	52.27%	6.82%	0.00%	40.91%	Anchor
	9.76%	39.02%	29.27%	21.95%	Reporting
	6.11%	23.33%	63.89%	6.67%	Reportage
	6.25%	18.75%	0.00%	75.00%	Graphics
	Overall Accuracy				59.07%

7. Conclusions

In this paper, a generic motion-based approach to semantic video analysis was presented. The proposed algorithm is based on the extraction of higher order statistics of the motion energy estimates and a new representation for providing local-level motion information to HMMs. Future work includes the investigation of corresponding algorithms for color/audio signal processing that will allow the integration of the proposed motion-based approach in a multi-modal video analysis scheme.

References

[1] S. Chang. The holy grail of content-based media analysis. *Multimedia, IEEE*, 9(2):6–10, 2002.

[2] G. Giannakis and M. Tsatsanis. Time-domain tests for Gaussianity and time-reversibility. *Signal Processing, IEEE Trans. on*, 42(12):3460–3472, 1994.

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

[4] J. Huang, Z. Liu, and Y. Wang. Joint scene classification and segmentation based on hidden Markov model. *Multimedia, IEEE Trans. on*, 7(3):538–550, 2005.

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

[6] 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.

[7] M. Proesmans, L. Van Gool, E. Pauwels, and A. Oosterlinck. Determination of Optical Flow and its Discontinuities using Non-Linear Diffusion. *Computer vision-ECCV'94: Third European Conf. on Computer Vision Stockholm, Sweden, May 2-6, 1994: Proceedings*, 1994.

[8] L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[9] D. Sadlier and N. O'Connor. Event Detection in Field Sports Video Using Audio–Visual Features and a Support Vector Machine. *CSVT, IEEE Trans. on*, 15(10):1225, 2005.

[10] B. K. Sinha. Detection of multivariate outliers in elliptically symmetric distributions. *The Annals of Statistics*, 12(4):1558–1565, Dec. 1984.

[11] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *PAMI, IEEE Trans. on*, 22(12):1349–1380, 2000.

[12] L. Xie, P. Xu, S. Chang, A. Divakaran, and H. Sun. Structure analysis of soccer video with domain knowledge and hidden Markov models. *Patt. Recog. Letters*, 25(7):767–775, 2004.