Detection of Multiple Subevents in Space and Time for Video Analysis

Alexia Briassouli*, Ioannis Kompatsiaris†,
*Corresponding author,Informatics and Telematics Institute, Thessaloniki, Greece
†Informatics and Telematics Institute, Thessaloniki, Greece

***Abstract*— This work presents a principled system based on statistical signal processing for the detection of regions of motion in video and the determination of time instants when there are changes in activities. Currently, the detection of temporal changes in video detects different shots instead of changes in activities. Areas of activity are first found by processing higher order statistics of illumination changes. Subsequently, change detection techniques are applied to sequentially estimated likelihood ratios of the illumination changes, to find changes in their distribution, that correspond to the beginnning or end of activity. Experiments with real videos demonstrate the accuracy of the temporal and spatial localization as well as their usefulness in practical applications.**

I. Introduction

One of the most challenging parts of designing a system for video analysis is the incorporation of intelligence into it, so that it can extract the time and location of events that are interesting, with minimal (or none) user intervention, and prior knowledge. Current systems often suffer from the drawback of being trained to detect specific events and objects, which limits their adaptability and applicability to different operating scenarios [1]. This work presents a principled system for detecting where and when new activities occur in a video. The system is based on statistical analysis of the illumination changes, rather than on prior knowledge or training. Thus, they can be applied in various applications, and can also be used to extract higher level knowledge, for example by using rules or knowledge structures, such as ontologies.

A. Previous Work, Motivation

There have been various approaches to the detection of events of interest in videos. One large category of approaches searches for specific spatio-temporal occurrences, connected in a particular order in time, referred to in [2] as "chronicles", which may be used to also draw conclusions about previous or future events in a video. In [1], regions of activity are extracted, , and objects of interest are then extracted, to understand the content of a scene and the events in it, based on a predefined scenario. In [3] characteristic areas of human activities are found by thresholding inter-frame differences in video. However, these methods only localize events in space, and not in time. The method we propose overcomes this limitation by using sequential likelihood ratio testing to find when new events occur. In [4], spatio-temporal clusters where events are expected to occur are extracted, and are modeled and grouped based on Gaussian mixture modeling (GMM) [5]. This has the advantage of not imposing any constraints on the object and region shapes or motions, but requires training, and the assumption that each spatio-temporal cluster remains according to a specific GMM model. As a consequence, these systems are not adaptable to different applications and entail a significant computational cost for each new problem.

B. Contributions

This work proposes a generally applicable system for automated detection, both in time and in space, of activities in video. The combination of its results with prior spatial knowledge about a scene, as in [6], [7] can lead to higher level knowledge about the events. Nevertheless, the approach is not tailored to a specific application, nor requires training, so it can be applied to a variety of input data. As a first stage, higher order statistics of inter-frame illumination variations are processed (Sec. II) to determine which pixels undergo motion during the entire video (Fig. 1(a)). Only these pixels are processed in the sequel, with the advantages:

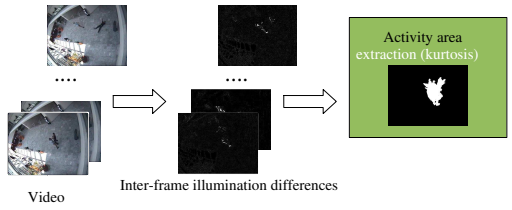
1) We only process pixels that are active, so there are fewer false alarms, i.e. much fewer static pixels are erroneously characterized as active.
2) Low computational cost: active pixels are much fewer than all the pixels in each frame.

Changes in the data [8], [9] are then found by sequential likelihood ratio testing on the illumination variations (Fig. 1(b)). The input data disrtibution is modeled empirically [10], thus increasing the system's flexibility and generality. Finally, activity areas corresponding to multiple sub-events are found by processing the higher order statistics of the illumination variations over the extracted times (frames) of change.

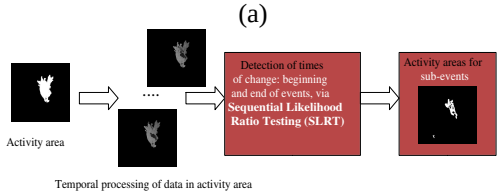
II. Kurtosis-based Spatial Localization of Activity Areas

Illumination variations are processed statistically to extract activity areas, i.e. binary masks that indicate which pixels move over the frames examined. Illumination variations are easily and accurately estimated by simple frame differencing in indoors videos, or videos filmed in controlled environments. For data with increased noise, illumination changes are better approximated by flow estimates [11]. For simplicity, we refer to either inter-frame illumination differences or optical flow measurements as "illumination changes" $d(x, y, t)$.

In practice, illumination changes are never perfectly accurate, so measurement noise is added to their estimates. Due to the large volume of video data and the independence of the noise sources, the Central Limit Theorem (CLT) holds,



(a)



(b)

Fig. 1. (a) Procedure for the extraction of the frames at the beginning and end of events, and the localization of pixels that move during each event.

so measurement noise is approximated by a Gaussian distribution[1] [12]. Illumination changes caused by motion deviate from the Gaussian approximation, since object motion is quite different from random measurement noise. This leads to the following two hypotheses for velocity estimates in frame k:

$$\begin{aligned} H_0 : v_k^0(\bar{r}) &= z_k(\bar{r}) \\ H_1 : v_k^1(\bar{r}) &= u_k(\bar{r}) + z_k(\bar{r}). \end{aligned} \quad (1)$$

H_0 expresses a velocity estimate at pixel $\bar{r}$, in frame k, introduced by measurement noise, and H_1 corresponds to the measurement of true motion $u_k(\bar{r})$, in the presence of noise $z_k(\bar{r})$ [9]. Thus, we can detect which velocity estimates correspond to a moving pixel by examining the non-gaussianity of the accumulated velocity estimates [13]. The non-gaussianity of random variable y can be tested using its kurtosis, defined as follows:

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

because the kurtosis of a Gaussian random variable is equal to zero. In practice, illumination variations in static pixels are not strictly Gaussian: they are often modeled as mixtures of Gaussians [14]. Nonetheless, the kurtosis in moving pixels is still significantly higher than in static ones, as it is proven to be very sensitive to outliers [15], [16], even under non-Gaussian noise [17].

A. Effect of additive noise on kurtosis

We consider the kurtosis of Gaussian data y, corrupted by additive noise v:

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

where $E[(y+v)^4] = E[y^4] + E[v^4] + 6\sigma_y^2\sigma_v^2$, so Eq. (3) becomes:

$$\begin{aligned} kurt(y+v) &= E[y^4] + E[v^4] + 6\sigma_y^2\sigma_v^2 - 3(\sigma_y^2 + \sigma_v^2)^2 \\ &= E[y^4] + E[v^4] - 3\sigma_y^4 - 3\sigma_v^4 \\ &= E[v^4] - 3\sigma_v^4 = kurt(v), \end{aligned} \quad (4)$$

[1]According to the CLT, when random variables result from the addition of a large number of independent noise samples (that follow any distribution), they can be modeled by a Gaussian distribution

where we took into account that the kurtosis of y is equal to zero, since y is a Gaussian random variable. Thus, we see that the kurtosis for non-Gaussian additive noise is, as expected, not equal to zero, but equal to the kurtosis of the additive noise v. If the additive noise is Gaussian, and we assume that y and v are independent, the kurtosis of $y+v$ becomes zero again, because the sum of independently distributed Gaussian random variables is also Gaussian [18].

III. Temporal Localization of Activities

We base the temporal localization of the beginning and end of activities on sequential statistical hypothesis testing [19], [20], which is used for change detection [9], [21]. Statistical change detection techniques are very well suited to the extraction of time instances where an activity begins and/or ends, since the distribution of a pixel's illumination changes is different when that pixel's motion changes. A significant advantage of these methods is that the test can be implemented as each new data sample arrives, in order to extract information about its distribution [9], [20], [19], thus allowing for real-time processing and quickest detection, so changes/events are detected with minimum delay.

A. Temporal Event Localization Algorithm

The input is a temporal sequence of illumination changes from frame 1 to t:

$$\bar{v}^t(\bar{r}) = [v_1(\bar{r}), v_2(\bar{r}), ..., v_t(\bar{r})]. \quad (5)$$

The illumination change at pixel $\bar{r}$, between frames $i-1$ and i, is $v_i(\bar{r}) \in V$. In order to determine whether a chang has occurred at an *unknown* time instant t, we formulate a binary hypothesis test: the data before the (unknown) change instant τ follows P_0, and after the change (H_1), it follows P_1.

$$\begin{aligned} H_0 : v_i(\bar{r}) &\sim P_0 \\ H_1 : v_i(\bar{r}) &\sim P_1. \end{aligned} \quad (6)$$

In the most general case, the data distributions before and after a change are not known, so they are empirically determined [22], [23].

B. Sequential Likelihood Ratio Testing - CUSUM Algorithm

In order to find a change in the data distribution from P_0 to P_1, the log-likelihood ratio is examined. For data samples from time instant 1 to t, we have:

$$T_1^t = \log \frac{\Pr(v_1(\bar{r}), ..., v_t(\bar{r})|P_1)}{\Pr(v_1(\bar{r}), ..., v_t(\bar{r})|P_0)}. \quad (7)$$

Assuming the data samples $v_i(\bar{r})$ are independently distributed, Eq. (7) becomes:

$$T_1^t = \sum_{i=1}^{t} \log \frac{(P_1[v_i(\bar{r})])}{(P_0[v_i(\bar{r})])} = T_1^{t-1} + \log \frac{(P_1[v_t(\bar{r})])}{(P_0[v_t(\bar{r})])}. \quad (8)$$

If a change occurs at frame t, the data after frame t is more likely to follow P_1 than P_0, so the ratio T_1^t is higher than a threshold η, determined according to Wald's approximation [9], [20] as follows:

$$\eta = \log((1-\beta)/\alpha), \quad (9)$$

where, α is a user-specified probability of false alarm

$$\alpha = \Pr(H_1|H_0) = \Pr(T^t \geq \eta|H_0), \quad (10)$$

and β is the probability of miss:

$$\beta = \Pr(H_0|H_1) = \Pr(T^t < \eta|H_1). \quad (11)$$

As Eq. (7) shows, a natural consequence of this test is that the decision for the change instant can be made online, as each new measurement arrives at time t.

C. Empirical Distribution Approximation

A challenging issue in all event detection problems is that the probability distributions before and after the change point are unknown. In practice, we can assume that P_0 is already known, or estimated empirically from the data, from the first τ_0 observations [22]. This implies the necessary assumptions that enough samples have been collected to provide a reliable distribution estimate (P_0), and that no change has occurred until τ_0. This may be a limitation in practice, since there might be changes before τ_0, however it is a necessary assumption, due to complete lack of any other source of information. The distribution P_1 after the change is approximated by incrementally updating the original estimate P_0. Thus, P_0 for the data in $\bar{v}^{\tau_0}(\bar{r}) = [v_1(\bar{r}), v_2(\bar{r}), ..., v_{\tau_0}(\bar{r})]$ is:

$$P_0[\bar{v}^{\tau_0}(\bar{r}) = j] = \frac{S_1^{\tau_0}[j] + \gamma}{w + \gamma \cdot |V|}, \quad (12)$$

where $S_1^{\tau_0}[j]$ is the frequency of data equal to $j \in V$ from frame 1 to τ_0:

$$S_1^{\tau}[j] = |\{i|v_i(\bar{r}) = j, 1 \leq i \leq \tau\}|. \quad (13)$$

The symbol $|\cdot|$ denotes the number of indexes i of the data that are equal to the value $j \in V$, so it corresponds to the domain size, and $\gamma = 0.5$ as in [22]. The distribution P_1 is based on incremental modifications of the original P_0 at each $t > \tau_0$:

$$P_1[\bar{v}^t(\bar{r}) = j] = \frac{S_1^t[j] + \gamma}{w + \gamma \cdot |V|}. \quad (14)$$

This estimate of P_1 has an inherent bias towards the distribution P_0 of previous samples. In order to avoid it, we window the data [24] to give a lower emphasis on older data samples. Since no prior knowledge is available about the data, the window size is determined empirically. We found that windows spanning 20 to 50 frames led to reasonable results for our experiments. Similarly, the threshold of Eq. (9) was empirically determined for a range of α and β that provided reasonable estimates of change (see Sec. IV).

1) Temporal Clustering for Localization of Activities: The proposed method for detecting changes indeed accurately localizes instants where changes occur in activity. In practice the estimates of change times may be close to each other, but not exactly the same, due to small numerical errors. This is easily overcome by clustering the estimated times of change: e.g. if new activity begin in spatially close pixels at times $38, 39, 41, 42$, its beginning is assigned to time instant 40 for all those pixels. Here, K-means is used for the clustering, but more advanced methods like Spectral Clustering could also be used. In practice, we empirically determined that a number of three clusters is sufficient for the videos examined here. In most cases, there are more clusters than events, so they are merged when the activity in them is of similar magnitude and direction. More details regarding the precise localization of clusters will be included in future research extensions of this work.

IV. Experiments

Experiments used well-known test sequences for surveillance applications, (from http://www-prima.imag.fr/PETS04/caviar data.html) and an outdoors traffic video. For initialization, we assumed no change in activity for the first 10 frames; this is realistic in the videos examined, as they contain more than 100 frames. Quantitative results also show the accuracy of these results when compared with manually generated ground truth in Table I.

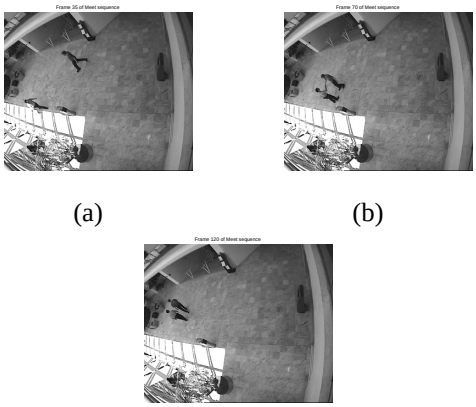
A. Meet sequence

We first examine a video of two people meeting and leaving together (Fig. 2). The activity area (Fig. 3(a)) has a characteristic shape, showing the men walking towards each other, and then leaving together. The illumination changes of the pixels in this area are processed to find at which frames events begin and/or end: there are 9539 pixels in the activity area, significantly reducing computations, since the 288×384 frames have 110592 pixels.

We compare Eq. (7) with Wald's threshold. For the threshold, we test values of $\alpha,\ \beta$ ranging from 0.01 to 0.04, with 0.01 increments, and choose $\alpha = 0.1$, $\beta = 0.08$. Fig. 3(b) shows the likelihood ratio values plotted against time, for all pixels in the activity area of Fig. 3(a). During the first frames, two groups of pixels area are activated, corresponding to the two people walking towards each other. After frame 90, there is only one group of active pixels, that corresponds to the people walking together. The activity areas of the two subevents are then estimated (Fig. 4): there is one subevent during frames $1-70$, since at frame 70 many pixels are de-activated. Between frames $70-90$ there is no significant activity, as the two men are standing and talking to each other and shaking hands, which can be seen in the resulting activity area (Fig. 4(b)). After frame 90, the pixels where the men walk as they leave the room are activated, leading to the activity area of Fig. 4(c).

B. Leave Box

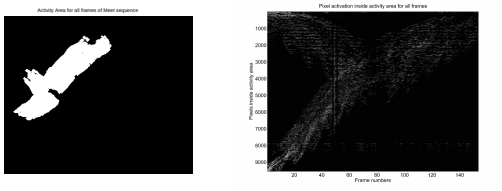
In this experiment,a woman enters a lobby, leaves a box, and then leaves (Fig. 5), resulting in the activity area of Fig. 6(a). Fig. 6(b) shows how the likelihood ratio values vary with time, as the woman moves over different active pixels: their values become low, near-zero, when the woman is not walking in those areas. Some of those pixels become active again in the last frames of the sequence, because the woman returns to that area of the room, after leaving the box. Thus, we see that the increases/decreases of the likelihood ratio values of Fig. 6(b), and the corresponding subevent activity areas of Fig. 7 agree with the actual events and human observation. Finally, it should be noted that the subevent activity areas of Fig. 7(a), (c) are similar, as the same pixels are activated when the woman walks before and after leaving the box. However, the direction of her motion in those two areas is opposite.



(a) (b)

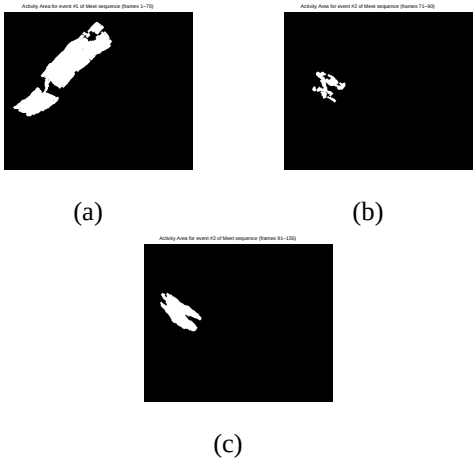
(c)

Fig. 2. Surveillance video of two men meeting. Frames (a) 35, (b) 70, (c) 120.



(a) (b)

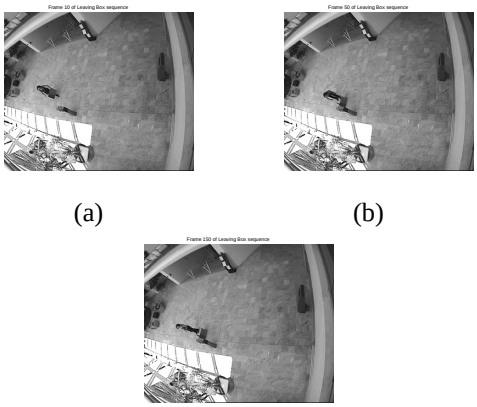
Fig. 3. (a) Activity area for all frames. (b) Likelihood ratio values for active pixels.



(a) (b)

(c)

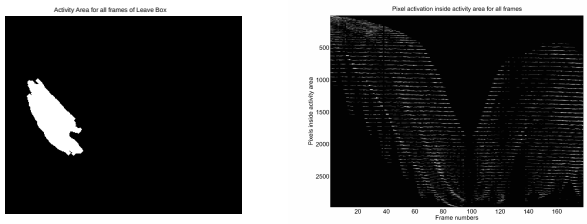
Fig. 4. Activity areas for subevents. (a) Men walking towards each other from opposite directions. (b) Men shaking hands. (c) Men leave room together.



(a) (b)

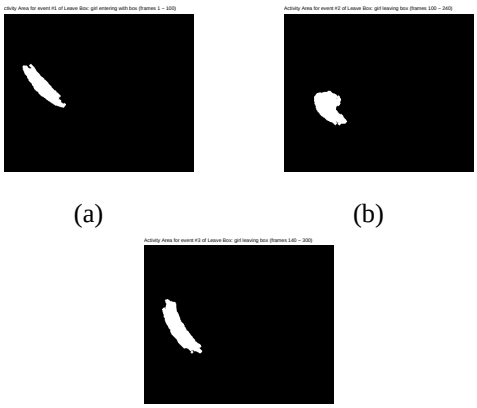
(c)

Fig. 5. Surveillance video of woman leaving a box in a room. (a) Frame 10, woman comes in with box. (b) Frame 50, woman leaves box. (c) Frame 150, woman leaves without the box.



(a) (b)

Fig. 6. (a) Activity area for all frames. (b) Likelihood ratio values for active pixels.



(a) (b)

(c)

Fig. 7. Activity areas for subevents. (a) Woman enters with box. (b) Woman leaves box. (c) Woman leaves without box.

C. Outdoors Taxi sequence

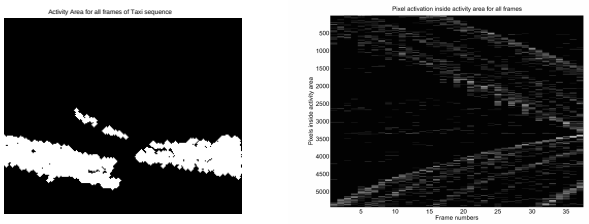
The outdoors Taxi sequence (Fig. 8) is used to examine the performance of the proposed method for a challenging outdoors video of poor quality. The activity area of Fig. 9(a) is derived using optical flow estimates instead of inter-frame illumination differences, because of the video's low resolution and the fact that it is shot outdoors. Fig. 9(b) shows how the likelihood ratio values change: two large pixel groups are activated, corresponding to the cars that are moving towards each other. A third, smaller group of pixels is also activated during the same frames, corresponding to the taxi turning. Fig. 9(b) shows that at the last video frames, the active pixels are in the same location, because the cars have met. This video contains only one event, as the cars are only moving until they meet, so the subevent activity areas coincide with the activity area of Fig. 9(a). Even for this poor quality data, the experimental results are accurate, as the statistical methods used are robust to noise, and are based on principled statistical

tests based on the actual data distributions, rather than ad-hoc thresholding or other heuristics.

(a) (b)

(c)

Fig. 8. Taxi sequence of cars approaching intersection and each other, and taxi turning.



(a) (b)

Fig. 9. (a) Activity area: pixels where cars approach each other, and taxi turns, are white. (b) Likelihood ratio values for pixels inside the activity area, over all video frames.

D. Quantitative Evaluation

In this section, we perform a quantitative evaluation of the proposed system's performance. Ground truth for the activity areas corresponding to the entire video and the subevents is extracted manually, by finding which pixels are active over the frames of interest. The ground truth for the beginning and end times of activities is also found manually, by observing at which frames some groups of pixels start and others stop to be active. Table I shows the relative absolute error, computed as the absolute difference between the ground truth and estimated activity area pixels, divided by the total number of pixels in the ground truth activity area. The same error, estimated for each subevent's activity area, is also derived. The mean of the subevents' relative errors is estimated and shown in the second line of Table I. Also, the mean of the absolute difference between the estimated times of change in activity and the corresponding ground truth for each video is shown in the last line of Table I. These results show that, as expected from the qualitative results of the previous sections, the proposed system gives reliable results for the activity areas and the times at which they occur.

TABLE I
MEAN ABSOLUTE ERROR OF ACTIVITY AREAS COMPARED WITH GROUND TRUTH.

Error—Video	Meet	Leave Box	Taxi
Activity Area	0.32%	0.21%	0.56%
Subevent Activity Area	0.36%	0.27%	0.56%
Time Instants of Change	2.4%	3.3%	4%

V. CONCLUSIONS, FUTURE WORK

A novel method for the spatiotemporal localization of multiple activities in video has been presented. The pixels which undergo displacement, and therefore correspond to event locations, are detected by processing the higher order statistics of accumulated inter-frame illumination differences. The detection of the start and end times of subevents is achieved via sequential empirical likelihood ratio testing. This is a significant contribution, as current surveillance techniques focus only on the spatial localization of activities, and not on the detection of the start or end of multiple subevents. Finally, fusion of the temporal localization results with the spatial processing method leads to characteristically shaped activity areas. Future extensions include the development of a complete ground truth set for subevents of interest, which would enable the determination of probabilities of false alarm and miss that are realistic in a wide range of surveillance applications. This will allow thorough experimentation with controlled (but meaningful) probabilities of false alarm and miss, for the development of surveillance systems with controlled sensitivity and reliability.

VI. ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Communitys Seventh Framework Programme FP7/2007-2013 under grant agreement n215453 - WeKnowIt, from FP6 under contract number 027685-MESH and 045547- VidiVideo.

REFERENCES

[1] Ziliani, F., Cavallaro, A.: Image analysis for video surveillance based on spatial regularization of a statistical change detection. In: Proc. 10th Int. Conf. on Image Analysis and Processing, Venice (1999) 1108–1111
[2] Ghallab, M.: On chronicles: Representation, on-line recognition and learning. In Aiello, D., Shapiro, eds.: Proc. Principles of Knowledge Representation and Reasoning. (Nov. 1996) 597–606
[3] Bobick, A.F., Davis, J.W.: The recognition of human movement using temporal templates. IEEE Transactions on Pattern Analysis and Machine Intelligence **23**(3) (March 2001) 257–267
[4] Belongie, S., Carson, C., Greenspan, H., Malik, J.: Color- and texture-based image segmentation using em and its application to content-based image retrieval. In: Proceedings of the 6th Int. Conference on Computer Vision, ICCV98. (1998)
[5] Redner, R.A., Walker, H.F.: Mixture densities, maximum likelihood and the em algorithm. SIAM Review **26**(2) (1984) 195–239
[6] Kwak, S., Bae, G., Kim, K., Byun, H.: Unusual event recognition for mobile alarm system. In: Computational Science ICCS 2007. (July 2007) 417–424
[7] Desurmont, X., Hayet, J.B., Delaigle, J.F., Piater, J., Macq, B.: TRIC-TRAC video dataset: Public HDTV synthetic soccer video sequences with ground truth. In: Workshop on Comp. Vision Based Analysis in Sport Env. (CVBASE). (2006)
[8] Zoua, C., Liu, Y., Qina, P., Wanga, Z.: Empirical likelihood ratio test for the change-point problem. Stats. and Probability Letters **77**(4) (Feb. 2007) 374–382

[9] Poor, H.V.: An Introduction to Signal Detection and Estimation. 2nd edn. Springer-Verlag, New York (1994)
[10] Einmahl, J., McKeague, I.: Empirical likelihood based hypothesis testing. Bernoulli 9 (2003) 267 – 290
[11] Duncan, J., Chou, T.: On the detection of motion and the computation of optical flow. IEEE Transactions on Pattern Analysis and Machine Intelligence **14**(3) (March 1992) 345–352
[12] Papoulis, A.: Probability, Random Variables, and Stochastic Processes. 2nd edn. McGraw-Hill, New York (1987)
[13] Giannakis, G., Tsatsanis, M.K.: Time-domain tests for Gaussianity and time-reversibility. IEEE Transactions on Signal Processing **42**(12) (Dec. 1994) 3460 – 3472
[14] Stauffer, C., Grimson, W.: Adaptive background mixture models for real-time tracking. In: Proc. IEEE Conf. Computer Vision Patt. Recog. (June 1999) 246–252
[15] Regazzoni, C.S., Sacchi, C., Teschioni, A., Giulini, S.: Higher-order-statistics-based sharpness evaluation of a generalized gaussian pdf model in impulsive noisy environments. In: Statistical Signal and Array Processing, 1998. Proceedings., Ninth IEEE SP Workshop on. (Sept. 1998) 411 – 414
[16] Nandi, A.: Robust estimation of 3rd-order cumulants in applications of higher-order statistics. Radar and Signal Proc., IEE Proceedings **140**(6) (Dec. 1993) 380–389
[17] Delaney, P.A.: Signal detection using third-order moments. Circuits Systems Signal Process **13**(4) (1994) 481–496
[18] Patel, J.K., Read, C.B.: Handbook of the Normal Distribution. Dekker, New York (1982)
[19] Basseville, M., Nikiforov, I.: Detection of Abrupt Changes: Theory and Application. Prentice-Hall, Inc., Englewood Cliffs, N.J. (1993)
[20] Wald, A.: Sequential Analysis. Dover Publications (2004)
[21] Nikiforov, I.: A generalized change detection problem. IEEE Transactions on Information Theory **41**(1) (Jan. 1995) 171 – 187
[22] Muthukrishnan, S., van den Berg, E., Wu, Y.: Sequential change detection on data streams. In: ICDM Workshop on Data Stream Mining and Management, Omaha NE (Oct. 2007)
[23] Shorack, G.R., Wellner, J.A.: Empirical Processes with Applications to Statistics. Wiley, New York (1986)
[24] Datar, M., Gionis, A., Indyk, P., R.Motwani: Maintaining stream statistics over sliding windows. SODA (2002) 635644