Human Activity Localization via Sequential Change Detection

Alexia Briassouli
Informatics and Telematics Institute, Centre for Research and Technology Hellas
abria@iti.gr

Ioannis Kompatsiaris
Informatics and Telematics Institute, Centre for Research and Technology Hellas
ikom@iti.gr

ABSTRACT

Today's rapid developments in digital media processing capabilities, and network speeds, make the dissemination of multimedia data extremely rapid and reliable, and have attracted significant research attention to video analysis, event detection, tracking and surveillance. In this work, a novel, generally applicable approach to the detection of human activity in video is presented. The areas of activity in the video are first detected via the accumulation and statistical processing of the motion vectors in all frames. The times (frames) at which events begin and end are defined as moments at which the statistical distribution of the motion vectors changes, for each pixel. These time instants are estimated in a novel manner, by applying sequential likelihood ratio testing on the motion vectors of the pixels that have been found to be active.

The proposed system provides a theoretically sound solution for the detection of temporal changes in the human (or other) activity in video, without resorting to use of prior knowledge, heuristics, or ad-hoc thresholds. Sequential detection techniques allow us to find the frames where events begin and end, but also allows to pre-define the desired probabilities of false alarm and miss for the system. This is entirely novel for the temporal localization of activities and events in the video processing literature. Finally, sequential change detection methods require the smallest number of samples to detect a change, so they ensure the fastest detection of events. Experiments are performed with real sequences, involving human activities, for varying probabilities of false alarm and miss. Comparison with ground truth results shows that, indeed, the proposed method leads to meaningful localization of events both in time and in space.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

Keywords

video surveillance, sequential detection, activity localization

1. INTRODUCTION

The large improvements in digital multimedia processing have enabled the fast and reliable production, processing and dissemination of large amounts of digital data. This can be very useful, but can easily become a time consuming and cumbersome problem, since huge amounts of data need to be processed in order to derive useful information and conclusions. For this reason, the automated extraction of higher level information, such as when and where activities occur, or who is in a video, using low-level image and video features (e.g. color, motion) has attracted significant attention. In this paper, we focus on the problem of accurately localizing the video frames, i.e. time instants, where activities begin and/or end at each pixel. Current work has focused on well-defined, specific problems, such as security [27], sports [16] (which are characterized by very specific rules), or in applications where an entire event is described as a pre-determined sequence of events [13]. The temporal segmentation of videos usually comprises of shot segmentation [1], which groups frames that have been filmed using the same camera. However, activity localization based on shot detection makes the underlying assumption that each shot contains a single activity, which is not the case in many practical situations.

This paper presents a novel algorithm capable of detecting the beginning and end times of actual activities in videos, without requiring prior knowledge, or being limited to any particular application. It is based on the statistics of the motion (and/or inter-frame illumination changes) in the video, and interprets changes in these distributions as indicators of a new event. Consequently, it is based only on the data extracted from each particular sequence. The generality of our approach allows it to be extended to more specific setups, by incorporating additional constraints (e.g. spatial information about the scene), for more specialized activity detection and recognition.

1.1 Overall System

We present an original, multi-stage system for the detection of events in video sequences. At the first stage, the pixels at which any activity occurs at any frame subsequence, are found. This is achieved by accumulating all video frames and then processing the inter-frame variations in illumination based on their statistics (Sec. 2). Once these pixels are found, a sequential likelihood ratio test (SLRT) method is designed, in order to detect at which frames an event begins and/or ends, in each pixel. The SLRT that is proposed can be based on either the modeling of inter-frame illumination variations with a known distribution (e.g. Gaussian), or by using empirical probability density functions. After the frames at which events begin and/or end are found, the areas of those particular events are extracted, by processing the spatial frame data for each detected "event subsequence".

2. ACTIVITY AREA EXTRACTION

In many practical applications involving human activities, such as surveillance or tracking, the videos are usually filmed from static cameras. Thus, the changes in illumination between successive video frames can be attributed to object motions, and hence activities. If the camera does undergo motion, it can be compensated for in a pre-processing, global motion estimation stage [11], so the proposed method can then be applied to the resulting sequence. In order to find the areas of activity, the inter-frame illumination differences first need to be extracted. In high quality indoors videos, e.g. for surveillance, we can directly use simple frame differences [5], as there is little noise in the background. If the more general case, e.g. if there is camera measurement noise or changes in the lighting of a scene, inter-frame illumination differences can be approximated by optical flow estimates [6], [3], for which numerous robust estimation methods exist [15], [4], [29]. In both cases, either by using simple frame differences, or by using optical flow estimates, the data is a set of inter-frame illumination variations, denoted by $d_k(\bar{r})$ at each pixel $\bar{r}$, between frames k and $k+1$.

2.1 Kurtosis-based activity area localization

In order to determine which pixels undergo motion and which are active, we first accumulate all inter-frame illumination variations at each pixel $\bar{r}$, over N frames, i.e. $d_k(\bar{r})$, for $1 \leq k < N$. They represent a moving or static pixel, i.e. one of the following hypotheses:

$$\begin{aligned} H_0 : d_k^0(\bar{r}) &= z_k(\bar{r}) \\ H_1 : d_k^1(\bar{r}) &= u_k(\bar{r}) + z_k(\bar{r}), \end{aligned} \tag{1}$$

where H_0 corresponds to a luminance difference caused only by noise $z_k(\bar{r})$, between frames k and $k+1$, and H_1 corresponds to the case where there is actual motion $u_k(\bar{r})$, as well as measurement noise $z_k(\bar{r})$. In order to distinguish between these two hypotheses, a model for the noise is needed. Measurement noise is unknown,but it is often modeled in the literature as Gaussian [14]. This is further supported, in the case of video processing, by the fact that a large number of video frames gives a large number of inter-frame differences, and hence a large number of random variables, which can then be approximated by a Gaussian distribution based on the Weak Law of Large Numbers [21].

Thus, in order to determine if a pixel is static or moving, we examine the Gaussianity of the accumulated inter-frame *illumination variations* $d_k(\bar{r})$. The classical measure of Gaussianity for a random variable y is its kurtosis, which is defined as:

$$\mathbf{kurt}(\mathbf{y}) = \mathbf{E}[\mathbf{y}^4] - \mathbf{3}(\mathbf{E}[\mathbf{y}^2])^2. \tag{2}$$

The fourth moment of a Gaussian random variable is $\mathbf{E}[\mathbf{y}^4] = \mathbf{3}(\mathbf{E}[\mathbf{y}^2])^2$, so is kurtosis is equal to zero. Consequently, the kurtosis of a time series of noise induced illumination differences, $d^0(\bar{r}) = [d_1^0(\bar{r}), ..., d_N^0(\bar{r})]$ should be equal to zero, indicating that a pixel $\bar{r}$ has not actually moved throughout the N video frames under examination.

It should be emphasized that, even if the measurement noise is not strictly Gaussian, kurtosis is a robust detector of outliers, as shown analytically in [10], [26], [19], [30]. Thus, it can be used to reliably detect at which pixels motion is present by detecting outliers in the inter-frame illumination variations, as verified by our experiments, in Sec. 5.

In [10] it is rigorously proven that the kurtosis is a robust detector of outliers in Gaussian noise, but that it can also detect them when they are embedded in non-Gaussian noise. In order to demonstrate this, we consider the case of non-Gaussian, zero-mean (without loss of generality, since the mean can be subtracted from our data set) additive noise v, added to a Gaussian random variable y:

$$\mathbf{kurt}(\mathbf{y}+\mathbf{v}) = \mathbf{E}[(\mathbf{y}+\mathbf{v})^4] - \mathbf{3}(\mathbf{E}[(\mathbf{y}+\mathbf{v})^2])^2. \tag{3}$$

The fourth order moment is then given by:

$$\begin{aligned} & \mathbf{E}[(\mathbf{y}+\mathbf{v})^4] = \mathbf{E}[(\mathbf{y}^2+\mathbf{v}^2+\mathbf{2yv})^2] \qquad (4) \\ = \; & \mathbf{E}[\mathbf{y}^4] + \mathbf{E}[\mathbf{v}^4] + \mathbf{6E}[\mathbf{y}^2\mathbf{v}^2] + \mathbf{4}(\mathbf{E}[\mathbf{yv}(\mathbf{y}^2+\mathbf{v}^2)] \\ = \; & \mathbf{E}[\mathbf{y}^4] + \mathbf{E}[\mathbf{v}^4] + \mathbf{6E}[\mathbf{y}^2]\mathbf{E}[\mathbf{v}^2] \\ + \; & \mathbf{4E}[\mathbf{y}]\mathbf{E}[\mathbf{v}]\mathbf{E}[(\mathbf{y}^2+\mathbf{v}^2)] = \mathbf{E}[\mathbf{y}^4] + \mathbf{E}[\mathbf{v}^4] + \mathbf{6}\sigma_\mathbf{y}^2\sigma_\mathbf{v}^2, \end{aligned}$$

where y is Gaussian, so its kurtosis is equal to zero, and we have made the assumption that y and v are independent from each other. Also:

$$\mathbf{E}[(\mathbf{y}+\mathbf{v})^2] = \mathbf{E}[\mathbf{y}^2] + \mathbf{E}[\mathbf{v}^2] + \mathbf{2E}[\mathbf{yv}] = \sigma_\mathbf{y}^2 + \sigma_\mathbf{v}^2, \tag{5}$$

so Eq. (3) becomes:

$$\begin{aligned} & \mathbf{kurt}(\mathbf{y}+\mathbf{v}) = \mathbf{E}[\mathbf{y}^4] + \mathbf{E}[\mathbf{v}^4] + \mathbf{6}\sigma_\mathbf{y}^2\sigma_\mathbf{v}^2 - \mathbf{3}(\sigma_\mathbf{y}^2 + \sigma_\mathbf{v}^2)^2 \\ = \; & \mathbf{E}[\mathbf{y}^4] + \mathbf{E}[\mathbf{v}^4] - \mathbf{3}\sigma_\mathbf{y}^4 - \mathbf{3}\sigma_\mathbf{v}^4 = \mathbf{E}[\mathbf{v}^4] - \mathbf{3}\sigma_\mathbf{v}^4 = \mathbf{kurt}(\mathbf{v}). \qquad (6) \end{aligned}$$

Again, we have taken into account that the kurtosis of y is equal to zero, since y is Gaussian, and 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 v was Gaussian, and 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 [12], [22].

2.2 Kurtosis estimates: Monte-Carlo testing for Gaussian, non-Gaussian data

We empirically demonstrate how the kurtosis of a Gaussian and noise-corrupted Gaussian random variable deviates from zero (Eq. (6)), as a function of the additive noise's variance σ_v^2, via 1000 Monte Carlo simulations. We generated 10^6 samples of a random variable y, following a zero-mean Gaussian distribution with variance equal to one, and estimated its kurtosis. In Fig. 1(a) the kurtosis of y is near zero (first point of the plot). Then, zero-mean Gaussian noise v, with increasing variance 0 to 1 by 0.01, was added to y, and the kurtosis of y and $y+v$ was estimated and plotted in the rest of Fig. 1. As expected, the values of the kurtosis remained low, in ± 0.01. In order to compare this with the kurtosis behavior under non-Gaussian noise, the same set of simulations was performed, but with the addition of Exponential noise

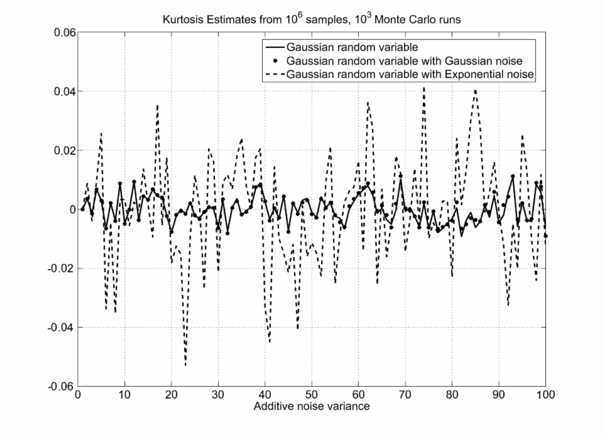

Figure 1: Kurtosis estimates for Gaussian random variables, Gaussian random variables with Gaussian noise and Gaussian random variables with Exponential noise.

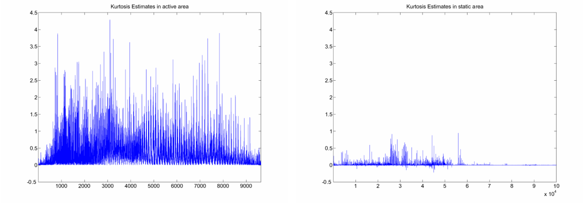

Figure 2: Kurtosis for "meet" sequence: (a) in activity area, (b) in static pixels.

v [21], [23], with the same energy as the previously used Gaussian noise (for fairness of comparison). Fig. 1 shows that, indeed, increasing non-Gaussian noise makes the kurtosis deviate from zero, with values in $(-0.006, 0.004)$. In our case, this is actually an advantage, as a large part of the additive noise is introduced to the measurement noise by the object velocity, under H_1 in Eq, (7).

2.3 Kurtosis estimates: real surveillance videos

In order to further justify the use of a Gaussian model for the measurement noise (i.e. for non-zero inter-frame illumination variations which are introduced by measurement, and not by motion), we conducted a series of experiments using the video sequence of Sec. 5.1. We obtained ground truth for the moving pixels, by manually delineating which area of the video frames undergoes motion. The inter-frame illumination variations of the static and active pixels over the entire video were then accumulated and the kurtosis of the resulting two sets ("static" and "active") of time series was estimated to examine their Gaussianity. The expectations $E[\cdot]$ in Eq. (2) and Eq. (3) are approximated by averaging over all moving and non-moving pixels.

Fig. 2 shows the results for the "meet" sequence in Sec. 5.1: the inter-frame illumination differences for active pixels lead to high kurtosis values, and deviate from a Gaussian distribution, whereas these variations over static pixels have much lower kurtosis values, even if the measurement noise distribution is not strictly Gaussian. This is because inter-frame illumination variations of moving pixels are outliers, when compared to the illumination variations of the static pixels, and higher order statistics, and specifically the kurtosis, are particularly sensitive to outliers [19], [30]. Here, the average of the kurtosis values of the moving pixels is equal to 0.3663, whereas the static pixels have a mean kurtosis equal to 0.0018.

2.4 Algorithm for activity area extraction

Thus, in order to find activity areas, containing the pixels that move throughout a video sequence, we (1) accumulate all inter-frame illumination differences (or flow estimates), (2) estimate their kurtosis and (3) extract a binary mask of active pixels from it by thresholding the higher values. Flow estimates and inter-frame illumination differences may contain outliers at the boundaries of moving objects [17], where the brightness constancy assumption is clearly violated. However, this does not introduce problems in our system, since the incorporation of outliers in the activity areas, may only cause a few additional pixels to be added to their border, which do not significantly affect its shape, nor the subsequent detection/classification of suspicious events or abnormal activities [7], [9]. This is verified in the experiments as well, in Sec. 5. Finally, it should also be noted that outlier pixels can be removed along with all static pixels in each frame, in a later stage, e.g. via color and/or texture processing, which refines the results of the optical flow processing [8].

The activity areas that are extracted in this manner correspond to the pixels that undergo motion throughout the video sequence, which may consist of more than one moving objects, in different locations. In that case, the activity area will actually include several different regions in each video frame. These regions allow us to estimate the number of independently moving objects in the video by extracting connected components from each activity area, and counting how many there are. This method can give a good estimate of the number of moving objects for the case of scenes that are not very crowded, but when the scenes contain many moving objects, e.g. people in a crowded train station, this estimate may become unreliable or even unfeasible. However, for surveillance applications of very crowded scenes, the precise estimation of the number of moving entities is extremely difficult, if not impossible, even for a human observer. The accurate estimation of moving entities in a crowded scene is thusly an ill-defined problem, and should be dealt with by different approaches [18], [25], [2].

3. DETECTION OF EVENT START AND END FRAMES

The detection of activity areas in video sequences is particularly useful, as it allows us to process only pixels which have undergone motion, in order to detect events. This has the advantage of reducing the computational burden significantly, since a much smaller amount of data is processed. Additionally, it increases the reliability of the overall system and decreases the number of "false alarms" in the event detection, since static pixels are not processed.

In order to detect at which time instants (frames) events begin or end at each pixel $\bar{r}$, we examine the time evolution of that pixel's inter-frame illumination variations $d_k(\bar{r})$, for $1 \leq k \leq N$. These variations form a dynamic phenomenon, which changes over time. A change in the distribution of the $d_k(\bar{r})$ indicates that an event has occurred, e.g. when a pixel has been static and starts to move at a frame k. Consequently we approach the problem of activity detection as a sequential likelihood ratio testing problem, where we find at which frame t the data does not follow its initial distribution.

For the time series in this problem, there are two hypotheses at each frame $1 \leq k \leq N$:

$$\begin{aligned} H_0 &: d_k^0(\bar{r}) \sim P_0 \\ H_1 &: d_k^1(\bar{r}) \sim P_t. \end{aligned} \tag{7}$$

P_0 is the initial data distribution, that we shall refer to as the *baseline distribution*, which is derived from the first w frames of the sequence (all $d_k(\bar{r}) where\ 1 \leq k \leq w$, follow P_0). This time w corresponds to the instant of the first change, which can be detected in an optimal manner processing all the available data simultaneously, by Change Detection techniques [20], as described below Eq. (9). In this work we focus on the sequential processing of the data, so w is pre-determined at a fixed value.

The video consists of N frames, but the change point(s) t is (are) unknown, so we do not know at which frame $w < t \leq N$ the distribution of $[d_t(\bar{r}), ..., d_N(\bar{r})]$ changes from P_0 to a different distribution, denoted P_t since we consider that change occurs at frame t. Consequently, the hypothesis H_1 is essentially a composite hypothesis, corresponding to "a change occurs after time w", and is expressed as:

$$H_1 = \bigcup_{w<t\leq N} H_t, \tag{8}$$

where each H_t corresponds to the hypothesis that a change occurred at time t, represented as:

$$H_t : d_k^1(\bar{r}) \sim P_t,\ t \leq k \leq N. \tag{9}$$

A change point can be detected via sequential likelihood ratio testing, i.e. by examining at each time instant k the value of the likelihood ratio:

$$L_k^N(\bar{r}) = \sum_{i=1}^{k} \log \frac{P_t(d_i(\bar{r}))}{P_0(d_i(\bar{r}))}, \tag{10}$$

where we have made the assumption that the inter-frame illumination variations $d_i(\bar{r})$ at each pixel $\bar{r}$ are independent from each other. This assumption is valid for the problem under examination, as under H_0, the inter-frame illumination variations are introduced by random noise, whose samples are independent of each other, whereas under H_1 (i.e. after a change) the illumination variation that originates from motion is, again, independent of the previous, noise-induced $d_i(\bar{r})$.

In Sequential Probability Likelihood Ratio Tests (SPRT), a decision about the data distribution is reached at the sample k^*, where the likelihood ratio $L_{k^*}^N(\bar{r})$ is below the lower threshold or above the higher threshold determined by Wald [24], [28]. These thresholds are determined based on user-defined probability of false alarm $\alpha = Pr(H_1|H_0) = Pr(L_k^N > \tau|H_0)$ and probability of miss $\beta = Pr(H_0|H_1) = Pr(L_k^N < \tau|H_1)$ as follows:

$$\tau_L = \log\left(\frac{\beta}{1-\alpha}\right), \quad \tau_H = \log\left(\frac{1-\beta}{\alpha}\right). \tag{11}$$

We then decide about the data distribution according to:

$$\begin{cases} L_k^N(\bar{r}) \leq \tau_L & d_k(\bar{r}) \sim H_0 \\ \tau_L < L_k^N(\bar{r}) < \tau_H & \text{more samples needed to decide} \\ L_k^N(\bar{r}) \geq \tau_H & d_k(\bar{r}) \sim H_1 \end{cases} \tag{12}$$

It has been proven [24] that this test enables us to decide about the data distribution using the smallest number of samples. Consequently, the time instant at which a pixel's activity changes, i.e. it either becomes active or inactive (its distribution changes from the baseline) will be detected quickly. Additionally, the thresholds are determined by predefined probabilities of false alarm and miss, which allow us to tune the sensitivity of the test.

The likelihood ratio values of (10) can also be used to determine (a priori) the first instant of change w. This requires estimating $L_k^N(\bar{r})$ for all time instants $1 \leq k \leq N$, and finding w such that $w = \text{argmax}_{1\leq k\leq N} L_k^N(\bar{r})$. The resulting time w is the time instant that has the highest probability of being a change point.

3.1 Gaussian Distribution Assumption

The sequential likelihood ration testing method presented in Sec. 3 is based on the assumption that the distributions P_0 and P_1 are already known. In this section, we consider the case where both data distributions are Gaussian, i.e. $P_0 \sim \mathcal{N}(\mu_0, \sigma_0)$ to $P_1 \sim \mathcal{N}(\mu_1, \sigma_1)$. This is in agreement with Sec. 2, which makes the assumption that the noise in the inter-frame illumination differences is approximately Gaussian. Then the test of Eq. (10) becomes:

$$L_k^N(\bar{r}) = \ln \frac{\sigma_0^2}{\sigma_1^2} + \sum_{i=1}^{k} \left[-\frac{(d_i(\bar{r}) - \mu_1)^2}{2\sigma_1^2} + \frac{(d_i(\bar{r}) - \mu_0)^2}{2\sigma_0^2} \right].$$

In this case, the form of the distributions is known, but their parameters still need to be estimated. For the baseline distribution P_0, μ_0, σ_0^2 can be approximated by estimating the mean and variance of the data until frame w, i.e. $[d_1(\bar{r}), ..., d_w(\bar{r})]$. The parameters of P_1 can then be estimated by incrementally updating the parameters of P_0, i.e. their mean and variance are re-estimated as new data $d_k(\bar{r}),\ k > w$ arrives. A possible drawback of this approach is that older data values, corresponding to P_0, are used to approximate P_1, making the test less sensitive to changes[1]. This is overcome in practice by setting a maximum memory length h for the data being examined, to ensure that P_1 remains up to date. In this work we focus only on the empirical data distribution, described in Sec. 3.2, as it provides a distribution model that is better tailored to the samples available.

3.2 Empirical Distribution Approximation

A similar approach can be followed to detect changes, without making any assumptions about the data distributions.

[1] P_1 does not differ enough from P_0 with each new data value, so changes are not detected as soon as they occur

This is the most challenging case for sequential likelihood ratio testing, namely Empirical Likelihood Ratio Testing, as it has no knowledge of either (1) the family of the distributions or (2) the time of change. However, it is also realistic, assuming a large enough number of samples is available (this is satisfied in video applications), since it uses the distribution of the data under examination.

In this work P_0 is estimated empirically using the "baseline data" $[d_1(\bar{r}), ..., d_w(\bar{r})]$, where $d_i(\bar{r}) \in \mathcal{D}$, $1 \leq i \leq N$. P_0 is found from the relative frequency with which each data value $\delta \in \mathcal{D}$ occurs:

$$P_0[\delta] = \frac{|\{i|d_i(\bar{r}) = \delta, 1 \leq i \leq w\}| + \gamma}{w + \gamma \cdot |\mathcal{D}|}. \qquad (13)$$

The quantity $|\{i|d_i(\bar{r}) = \delta, 1 \leq i \leq w\}|$ shows at how many time instants i the value $\delta \in \mathcal{D}$ occurs, γ is a smoothing parameter, usually set to 0.5 and $|\mathcal{D}|$ is the domain size. P_1 is estimated via incremental updates of the empirical distribution:

$$P_1[\delta] = \frac{|\{i|d_i(\bar{r}) = \delta, k - h + 1 < i \leq k\}| + \gamma}{h + \gamma \cdot |\mathcal{D}|}. \qquad (14)$$

The only difference from Eq. (13) is that the data used is $[d_{k-h+1}(\bar{r}), ..., d_k(\bar{r})]$ at each $k : w < k \leq N$, where h denotes the memory length, used to remain up to date.

4. CLUSTERING OF EVENT TIMES

In Sec. 3, we presented a theoretically sound method for finding instants where activities may begin or end, based on changes in the distribution of inter-frame illumination differences. However, in order to obtain practically useful detection results, we need to further process the extracted change points. In a real video, it is possible that several pixels in each activity area, e.g. pixels corresponding to a moving human's leg, are activated at time instants that are close to each other, but do not coincide. However, since those frames are close to each other, the changes in those pixel distributions should be assigned to the same activity (in this example, leg motion). For this reason, once the frames where changes occur in a video are estimated, we cluster them, to decrease the number of redundant or false alarm events.

An issue that remains, and is very difficult to resolve in an objective manner, is the number of clusters that should be used. The number of clusters will essentially determine the number of activities that are detected as well, however this depends on the definition of "activity", which remains subjective. A realistic solution with practical applicability is to determine the number of human activities experimentally, in the context of the application at hand. In our experiments, we found that events meaningful to humans are detected when approximately three clusters are used to group the extracted change points. This was determined after performing experiments with all the videos of Sec. 5, as well as twenty other similar indoors surveillance videos. The clustering method used in our experiments was K-means, however other methods, like mean shift mode seeking or spectral clustering can also be used.

5. EXPERIMENTS

Experiments were performed with real surveillance videos from the well-known CAVIAR-PETS test sets, which consist of indoors surveillance scenes, with scenarios like people walking and meeting, fighting, and also outdoors videos of people performing activities like skipping, walking towards each other. We first find the activity area by processing all the video frames at once, as in Sec. 2. Then, inter-frame illumination differences of the active pixels are estimated and SLRT is applied (Sec. 3) to find at which frames events begin and/or end. Testing is done for a large range of probabilities of false alarm and miss, to examine the sensitivity of the system to these parameters, as well as to find a range of these values that gives meaningful results in practical applications, but one characteristic pair is shown here for reasons of space. Finally, the activity area corresponding to each subevent is extracted, and examined by human observers (since there is no other form of ground truth for these applications), to determine if it corresponds to a meaningful event. As these results show, the proposed method is successful in segmentation of the video into activity-related segments, rather than video shots. This is a more meaningful result for high-level processing activities such as human motion or even detection, since a video shot is likely to contain more than one activities.

5.1 People Walking and Meeting

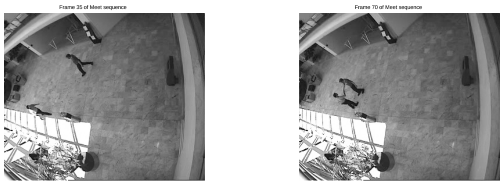

Figure 3: Meet sequence, frames 35, 70.

A sequence of two people walking towards each other, meeting, and then leaving together, is examined. Three characteristic frames, when they approach each other, meet, and leave together, are shown in Fig. 3. The activity area resulting from processing the higher order statistics of all inter-frame differences is shown in Fig. 4(a). The trajectories of the people approaching each other, and the common trajectory when they leave together, can be seen in this binary mask. SLRT (Sec. 3.2) determines time instants of change. The corresponding empirical likelihood ratio is shown in Fig. 4(b), where it can be seen that its values change near frames 70 and 90. The data was windowed using $h = 10$ to ensure that data more recent than 10 frames before the current one is being used.We set the probability of false alarm

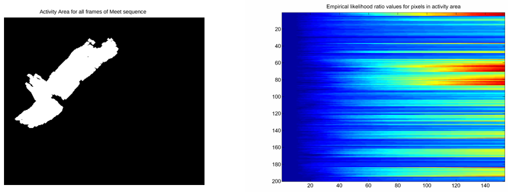

Figure 4: (a) Activity area, (b) SLRT for active pixels, over time.

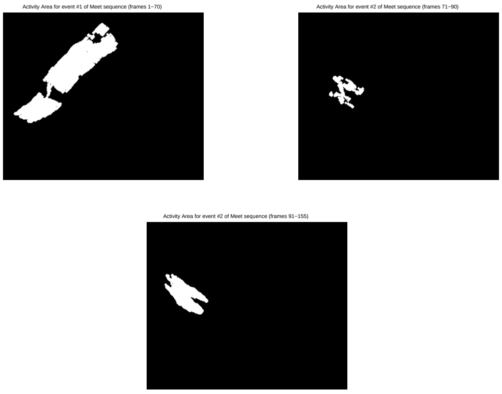
Activity Area for event #1 of Meet sequence (frames 1−70)
Activity Area for event #2 of Meet sequence (frames 71−90)
Activity Area for event #2 of Meet sequence (frames 91−155)

Figure 5: Activity areas: (a) Frames 1-70, people walking towards each other. (b) Frames 71-90, meet. (c) Frames 91-155, leave together.

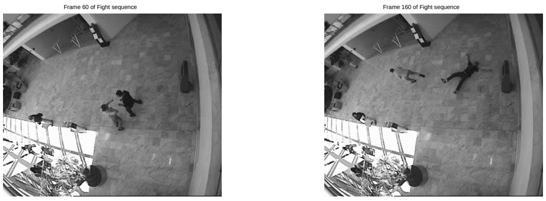

Figure 6: Fight sequence, frames 50, 160.

equal to 10^{-3} and the probability of miss equal to 10^{-4}, and after applying clustering, we indeed found the resulting frame subsequences are from frames $1-70$, $71-90$ and $91-155$. We then extract the activity areas for these subsequences, and the corresponding activities. As Fig. 5 shows, these areas are indeed representative of the activities taking place. In Fig. 5(a) the activity area is blob-like, when the people approach each other. In Fig. 5(b) the handshake between the two people can be clearly seen, whereas Fig. 5(c) shows the trajectory of them leaving together.

5.2 Fight

In this experiment, a video containing two people that walk towards each other and have a fight is examined. The "actors" first walk towards each other, during the fight they move in a circle, until one of them falls down, and then the other "actor" runs away (Fig. 6). The activity area extracted for the entire video sequence (180 frames) is shown in Fig. 7(a). In this mask, one can see the signature of the people walking towards each other, as well as a blob-like shape in the pixels where the fight took place. Sequential empirical likelihood ratio testing (Sec. 3.2) is applied to the data in order to determine the time instants of change. To determine the Wald threshold, we set the probabilities of false alarm and miss equal to 10^{-3}. The data is windowed using $h = 20$ to ensure that data more recent than 20 frames before the current one is being used. Fig. 7(b) shows the likelihood ratio values for the 4000^{th} pixel in the activity area, with the lower and upper Wald thresholds. From the figure, it is evident that after frame 6 we can decide that the data until then follows the initial distribution H_0, whereas at frame 52 the test exceeds the upper threshold, and the data follows H_1, i.e. there is activity. In this manner, the beginning and end times of all activities in the video are estimated, leading to "activity subsequences" corresponding to frames $1-60$, $61-140$ and $141-180$. The resulting activity areas for each subsequences, shown in Fig. 8, are indeed representative of the activities taking place. In Fig. 8(a), (b), the activity areas are blob-like, corresponding to the people approaching each other, fighting, and one man falling down. On the other hand, Fig. 8(c) has a linear shape, which is characteristic of a person walking or running, and, indeed, this subsequence corresponds to the one person running away.

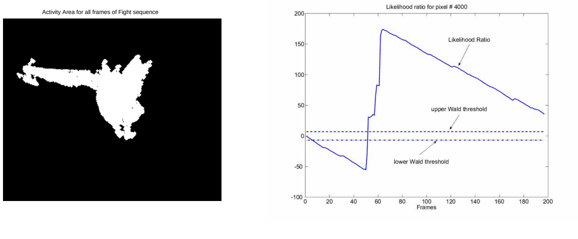

Figure 7: Fight sequence. (a) Activity Area. (b) SLRT.

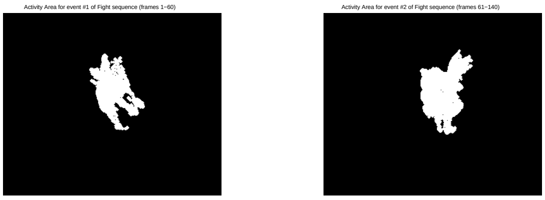
Activity Area for event #1 of Fight sequence (frames 1−60)
Activity Area for event #2 of Fight sequence (frames 61−140)

Figure 8: Activity areas: (a) Frames 1-60, people walking towards each other, fighting. (b) Frames 61-140, fight, fall down.

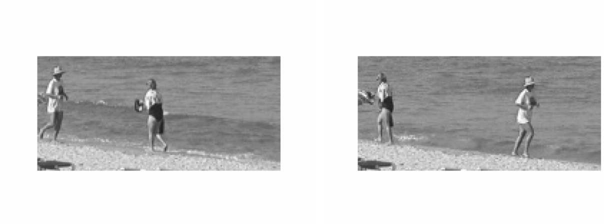

Figure 9: Ladies on beach: Frames 40, 70.

5.3 Ladies on beach

An outdoors video, of two ladies on a beach, one walking and one running, that cross each others' path, is examined (Fig. 9). The activity area of Fig. 10(a) shows in which area they were walking or running. After applying the SLRT to these pixels, for $P_{fa} = P_{miss} = 10^{-3}$, we obtain the log-likelihood ratios for all active pixels. A characteristic case is shown in Fig. 10(b), for the 2500^{th} pixel: we see that from

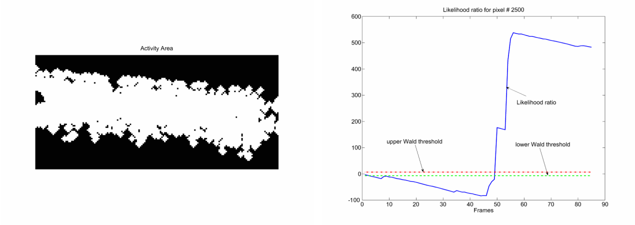

Figure 10: Ladies on beach: (a) Activity Area. (b) SLRT.

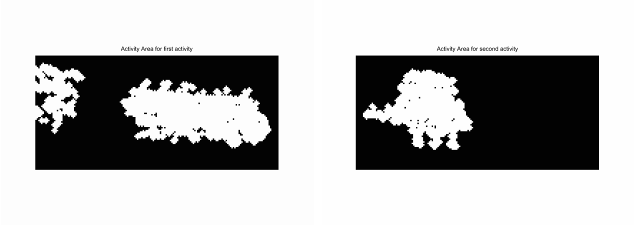
Activity Area for first activity
Activity Area for second activity

Figure 11: Activity areas: (a) before crossing, (b) during crossing.

frames $1 - 48$ the data followed P_0, for frames $48 - 54$ there is a change from H_0 ("no-decision" region until frame 55), and after frame 55 there is a new activity, corresponding to H_1. The corresponding activity areas, shown in Fig. 11, consist of (a) the subsequence where the two ladies approach each other, but have not met yet, (b) the frames where they cross paths, (c) the subsequence where they have crossed and continue walking/running in opposite directions.

6. CONCLUSIONS

A novel, generally applicable approach for the detection of the beginning and end of events in video sequences is presented. It is based on the empirical statistical modeling of inter-frame illumination variations and/or motion or flow vectors. The data is first processed using higher order statistics to extract areas of activity. The next stages process only pixels in these areas, in order to reduce computational cost and increase reliability. Each pixel inside an activity area is tracked over time, and its empirical distribution is estimated. Sequential empirical likelihood ratio testing is used to detect changes in this distribution, and map the corresponding frame numbers to the beginning or end of events. This allows a theoretically sound basis for temporal event localization, which can be extended to any practical application, as it does not involve any application-dependent heuristics. Another advantage of this approach is that, by its nature, sequential likelihood ratio testing allows us to predetermine probabilities of false alarm or miss. This adds another degree of flexibility to the system, as it allows us to control its sensitivity in detecting events depending on the needs of each application. Finally, sequential change detection methods require the smallest number of samples, to detect a change, so they ensure the fastest detection of events, and low computational cost. Experiments with real sequences show that the proposed method can lead to meaningful localization of events both in time and in space. Future work includes the further processing and combination of the extracted event times and activity area shapes, for the extraction of higher level semantics for videos. Since there is often no global ground truth for the beginning and/or end of events in the videos, future work will also focus on determining ground truth events for common applications.

7. REFERENCES

[1] H. A. Shot-boundary detection: unraveled and resolved? *IEEE Transactions on Circuits and Systems for Video Technology*, 12(2):90–105, 2002.

[2] E. Andrade, S. Blunsden, and R. Fisher. Hidden markov models for optical flow analysis in crowds. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, pages 460–463, 2006.

[3] J. Barron and R. Eagleson. Recursive estimation of time-varying motion and structure parameters. *Pattern Recognition*, 29(5):797–818, Dec. 1996.

[4] M. J. Black and P. Anandan. A framework for the robust estimation of optical flow. In *Proc. IEEE 4th Int. Conf. Computer Vision*, pages 231–236, May 1993.

[5] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):257–267, March 2001.

[6] J. Bouguet. Image sequence enhancement using multiple motions analysis. In *Intel Corporation. Microprocessor Research Labs.*

[7] A. Briassouli, S. Dasiopoulou, and I. Kompatsiaris. Ontology-based trajectory analysis for semantic event detection. In *Proceedings of the 1st IEEE International Conference on Semantic Computing, ICSC 2007*, Irvine, California, Sept. 2007.

[8] A. Briassouli, V. Mezaris, and I. Kompatsiaris. Color aided motion-segmentation and object tracking for video sequences semantic analysis. *International Journal of Imaging Systems and Technology (IJIST), Special Issue on Applied Color Image Processing*, 17(3):174–189, 2007.

[9] A. Briassouli, V. Mezaris, and I. Kompatsiaris. Video segmentation and semantics extraction from the fusion of motion and color information. In *Proceedings of the 2007 IEEE International Conference on Image Processing, ICIP 2007*, pages 2014–2017, San Antonio, Texas, Sept. 2007.

[10] P. A. Delaney. Signal detection using third-order moments. *Circuits Systems Signal Process*, 13(4):481–496, 1994.

[11] F. Dufaux and J. Konrad. Efficient, robust, and fast global motion estimation for video coding. *IEEE Transactions on Image Processing*, 9(3):497Ű501, March 2000.

[12] W. Feller. *An Introduction to Probability Theory and its Applications*, volume 1. John Wiley & Sons, New York, 1966.

[13] M. Ghallab. On chronicles: Representation, on-line recognition and learning. In D. Aiello and Shapiro, editors, *Proc. Principles of Knowledge Representation and Reasoning*, pages 597–606, Nov. 1996.

[14] R. Gonzalez and R. Woods. *Digital Image Processing.* Prentice Hall, New Jersey, 2002.

[15] L. Jacobson and H. Wechsler. Derivation of optical

flow using a spatiotemporal-frequency approach. *Computer Vision, Graphics and Image Processing*, 38:29–65, 1987.

[16] S. Kwak, G. Bae, K. Kim, and H. Byun. Unusual event recognition for mobile alarm system. In *Computational Science Ű ICCS 2007*, pages 417–424, July 2007.

[17] J. L.Barron, D. J. Fleet, S. S. Beauchemin, and T. A. Burkitt. Performance of optical flow techniques. In *1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 236–242, June 1992.

[18] A. Marana, M. Cavenaghi, R. Ulson, and F. Drumond. Real-time crowd density estimation using images. In *ISVC05*, pages 355–362, 2005.

[19] A. Nandi. Robust estimation of third-order cumulants in applications ofhigher-order statistics. *Radar and Signal Processing, IEE Proceedings*, 140(6):380–389, Dec. 1993.

[20] I. Nikiforov. A generalized change detection problem. *IEEE Transactions on Information Theory*, 41(1):171 – 187, Jan. 1995.

[21] A. Papoulis. *Probability, Random Variables, and Stochastic Processes.* McGraw-Hill, New York, 2nd edition, 1987.

[22] J. K. Patel and C. B. Read. *Handbook of the Normal Distribution.* Dekker, New York, 1982.

[23] P. Z. Peebles. *Probability, Random Variables and Random Signal Principles.* McGraw-Hill Inc, Boston, 2001.

[24] H. V. Poor. *An Introduction to Signal Detection and Estimation.* Springer-Verlag, New York, 2nd edition, 1994.

[25] H. Rahmalan, M. Nixon, and J. Carter. On crowd density estimation for surveillance. In *Crime and Security, 2006. The Institution of Engineering and Technology Conference on*, pages 540 – 545, June 2006.

[26] C. S. Regazzoni, C. Sacchi, A. Teschioni, and S. Giulini. 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*, pages 411 – 414, Sept. 1998.

[27] E. Stringa and C. S. Regazzoni. Content-based retrieval and real time detection from video sequences acquired by surveillance systems. In *Image Processing, 1998. ICIP 98. Proceedings. 1998 International Conference on*, volume 3, pages 138 – 142, Oct. 1998.

[28] A. Wald. *Sequential Analysis.* Dover Publications, 2004.

[29] J. Weijer and T. Gevers. Robust optical flow from photometric invariants. In *Proc. IEEE International Conference on Image Processing, 2004*, volume 3, pages 1835–1838, Oct. 2004.

[30] M. Welling. Robust higher order statistics. In *Tenth International Workshop on Artificial Intelligence and Statistics*, pages 405–412, Barbados, Jan. 2005.