STATISTICAL PROCESSING OF VIDEO FOR DETECTION OF EVENTS IN SPACE AND TIME

Alexia Briassouli, Ioannis Kompatsiaris

Informatics and Telematics Institute (ITI)
abria@iti.gr, ikom@iti.gr

ABSTRACT

Recently, advanced video processing systems have been developed for numerous applications, such as surveillance, tracking, monitoring, object and event detection. The large size of this data and the wide range of detectable events or objects render the automation of various video processing procedures necessary. The proposed system first finds the pixels where activities occur by processing the higher order statistics of luminance changes between accumulated video frames. Once these pixels are extracted, sequential likelihood ratio based change detection techniques are applied to find at which frames activities occur. The resulting system is novel, as it can localize actions both in space and in time, in a theoretically sound manner. Testing takes place with real video sequences, to demonstrate its effectiveness.

Index Terms— activity, event detection, temporal localization, sequential change detection, statistical processing

1. INTRODUCTION

The large advancements in digital multimedia processing make the automation of video systems particularly important. Fields like surveillance, monitoring and security have received considerable attention lately, and many methods have been developed for the detection of motion, activities, objects and events in video and other multimodal content. However, the precise localization of events in space and time is often related to particular applications, or requires heuristics in order to operate in a reliable manner.

In this work, we present a novel approach for the detection of pixels where activity occurs in video, and the extraction of the frames when changes occur in the video, i.e. where activities begin and/or end. The pixels where activity occurs form a binary activity mask, which is equal to one in the active pixels, and zero elsewhere. This mask is extracted by processing the higher order statistics of illumination differences between successive frames. Sequential change detection techniques are then applied to the luminance variations between video frames, in order to find at which frames changes occur. The current literature has not combined these statistical techniques in order to localize events, actions in space and time [1]. Also, unlike much of the current literature [2], [3], our method is independent of the kind of video, as it applies general statistical detection methods.

Section 2 presents a method for estimating the binary activity mask in a video. In Section 3, the statistical change detection approach for finding at which frames there is a change in the data distribution is presented. The results of Sec. 2 are used in Sec. 3 in order to reduce the computational cost and develop a more robust system. Experimental results are presented in Sec. 4 and conclusions and plans for future work are drawn in Sec. 5.

2. HIGHER ORDER STATISTICS FOR ACTIVITY MASK

Activity extraction in video is closely related to motion estimation, for which numerous methods have been developed. We consider that robust flow estimation methods [4], [5], [6] reliably extract inter-frame illumination variations used by our system. For high quality video, a lower computational cost can be achieved by using simple inter-frame luminance differences, instead of flow estimates.

Variations of luminance values originate either from pixel motion, or from measurement noise. We accumulate luminance variations over all frames and process their higher order statistics to detect pixels that undergo motion, and map them to a binary "activity mask". Differences in frame illumination between frames k and $k+1$, at pixel $\bar{r}$, are $d_k(\bar{r}),\ 1 \leq k < N$, for a video with N frames. The separation of active and static pixels corresponds to the following hypothesis test:

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

where $v_k(\bar{r})$ is a change in luminance introduced by motion, and $z_k(\bar{r})$ represents measurement noise. The accumulation of $N-1$ such differences over N video frames forms the random time series $D^i(\bar{r}) = [d_1^i(\bar{r}), ..., d_N^i(\bar{r})],\ i \in \{0, 1\}$. Under H_0, $D^0(\bar{r})$ only contains measurement noise $z_k(\bar{r})$, which is often modeled as Gaussian in the literature [7]. $D^1(\bar{r})$ follows an unknown distribution, due to the addition of unknown velocities $v_k(\bar{r}),\ 1 \leq k < N$, so testing the Gaussianity of

Fig. 1. Walking sequence. (a) Frame 1. (b) Activity Area.

$D^i(\bar{r})$ indicates if pixel $\bar{r}$ has undergone motion. This test would only fail if $D^1(\bar{r})$ also follows a Gaussian distribution, which is not realistic, since that would require that the pixel motion over frames 1 to N is precisely Gaussian as well.

The classic measure of a random variable's Gaussianity is its kurtosis [8]:

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

A Gaussian random variable y has $\mathbf{E}[\mathbf{y^4}] = \mathbf{3}(\mathbf{E}[\mathbf{y^2}])^\mathbf{2}$, so is kurtosis is equal to zero. Consequently, the kurtosis of $D^0(\bar{r})$ should be zero, indicating that a pixel $\bar{r}$ has not actually moved over the N video frames under examination. Even if the measurement noise is not strictly Gaussian, the kurtosis is a robust detector of outliers [9], [10], something which is also verified by our experiments (Sec. 4).

2.1. Algorithm for activity mask

In order to extract activity masks: (1) we form the time series $D^i(\bar{r})$ for each pixel $\bar{r}$, (2) we estimate the kurtosis of $D^i(\bar{r})$, $k_{D^i}(\bar{r})$, (3) we threshold $k_{D^i}(\bar{r})$ and set it equal to 1 if its value exceeds a (non ad-hoc) pre-determined threshold. Thresholded kurtosis values are accumulated for all pixels $\bar{r}$ to form the binary activity mask. Fig. 1 shows a frame from a video of a person walking and the resulting activity mask. Only the areas where there was motion are non-zero: a white "line" in the middle where the person crossed the room and a small blob on the bottom left where someone was moving locally. Note that the motion in the bottom left of the frame is not easily discernible by human observers, whereas the proposed method easily detects it. In practice, the kurtosis at static pixels has significantly lower values than at moving pixels, but is not exactly zero. Thus, we consider that the pixels whose kurtosis is above 10% of the mean value of the kurtosis values of all pixels, are active. Thus the resulting threshold adapts to the sequence under examination. The value of 10% was chosen after testing with the videos used in the experiments and 50 other videos, of similar scenes (e.g. surveillance), but also of different kinds, e.g. outdoors sports videos.

3. SEQUENTIAL LIKELIHOOD CHANGE TESTING FOR ACTIVITY DETECTION

A central contribution of this work is the development of a theoretically sound approach for the detection of the beginning or ending of activities in video. The data used is the same as in Sec. 2, i.e. the time series formed by the luminance variations between successive frames. We examine only the pixels corresponding to H_1, i.e. $D^1(\bar{r})$, since only those pixels undergo motion; this reduces the computational cost significantly. For example, in Fig. 1, the pixels in the activity mask are 3663, while there are 106929 static pixels (each frame is 288×384, i.e. has a total of 110592 pixels). This also increases the system's reliability and robustness, as there are no false alarms caused by detecting changes in pixels that are actually static.

Sequential likelihood ratio testing [11] is well suited suited for the detection of new activities in each pixel $\bar{r}$ using the smallest amount of data needed, as it has been proven to achieve quickest detection [11]. The test compares:

$$\begin{aligned} H_0 : D_k^0(\bar{r}) &\sim P_0 \\ H_1 : D_k^1(\bar{r}) &\sim P_1, \end{aligned} \quad (3)$$

where $P_i(D_k^1(\bar{r})) = P_i(d_1^i(\bar{r}), ..., d_k^i(\bar{r}))$, $i \in \{0, 1\}$. P_0 is a "baseline distribution", for data that does not undergo change, and P_1 corresponds to data that undergoes change. $d_n^i(\bar{r})$ can be assumed to be independently identically distributed (i.i.d.) under H_0, because they are equal to i.i.d. measurement noise. Under H_1, inter-frame velocity estimates added to the i.i.d. noise could introduce a degree of dependence, but cannot be modeled from before, since the motion is unknown. The i.i.d. assumption under H_1 is necessary, even if it does not always precisely hold, since no additional knowledge is available, and is shown to still provide reliable detection results. Then, the log-likelihood ratio is:

$$l_k(\bar{r}) = \ln \frac{\prod_{n=1}^k P_1(d_n^1(\bar{r})}{\prod_{n=1}^k P_0(d_n^1(\bar{r})} = \sum_{n=1}^k \ln \frac{P_1(d_n^1(\bar{r})}{P_0(d_n^1(\bar{r})}. \quad (4)$$

For initialization, we consider that there is no change in the first w frames of the sequence, so $D_w^1(\bar{r})$ follows P_0. For practical purposes, we set $w = 0.1N$, where N is the length of the data sequence under examination, a quantity which was experimentally shown to provide reliable change detection.

3.1. Gaussian Distribution Assumption

The distributions P_0, P_1 correspond to the data $D_k^1(\bar{r})$ in the activity masks, which in turn corresponds to hypothesis H_1 (i.e. active pixels) of Eq. (1). In Sec. 2 we based the activity mask extraction on the assumption of Gaussian measurement noise $z_k(\bar{r})$, so we also use the Gaussian model for P_0 and P_1. P_0 corresponds to the first w video frames, so its mean

μ_0 and variance σ_0^2 of P_0 can be easily estimated from the data $D_w^1(\bar{r})$.

For P_1, μ_1 and σ_1^2 can be estimated by incrementally updating μ_0 and σ_0^2, i.e. re-estimating them as new data $d_k^1(\bar{r}),\ k > w$ arrives. A possible drawback is that older data values, corresponding to P_0, are used to approximate P_1, making the test less sensitive to changes[1]. In order to overcome this problem, and ensure that P_1 remains up to date, we window the data: the parameters of P_1 are estimated at time k from the past h samples $d_n^1(\bar{r}),\ k-h<n<k$. Then:

$$l_k(D_k^1(\bar{r})) = \ln\frac{\sigma_0^2}{\sigma_1^2} + \sum_{n=1}^{k}\left[-\frac{(d_n^1(\bar{r})-\mu_1)^2}{2\sigma_1^2} + \frac{(d_n^1(\bar{r})-\mu_0)^2}{2\sigma_0^2}\right].$$

3.2. Sequential Change Detection

The sequential test is defined as:

$$\begin{array}{ll} l_k(D_k^1(\bar{r})) \geq \pi_U & H_1 \\ \pi_L < l_k(D_k^1(\bar{r})) < \pi_U & \text{obtain new sample} \\ l_k(D_k^1(\bar{r})) \leq \pi_L & H_0 \end{array} \quad (5)$$

This enables us to determine at which frame a change has occurred without using all data, allowing fast detection of events. The optimal estimation of $pi_U,\ \pi_L$ can be very difficult in practice [11].We determined them empirically as equal to 25% below and above the mean of $l_k(\bar{r})$ for π_L, π_U respectively. Sec. 4 shows that this led to the accurate detection of changes in the distribution of the luminance variation, and consequently the beginning and ending of new events.

4. EXPERIMENTS

In this section we present experimental results of the proposed method for real videos, used in surveillance applications.

Walking Sequence A video of a person walking (Fig. 1(a)) diagonally across a room was processed in this experiment. There is also small motion on the bottom left side of the video frame, where another person is moving, as seen in the activity mask in Fig. 1(b). Fig. 2(a) shows interframe illumination variations of the pixels in the activity mask. It can be seen that they are active during different subsequences of the video. We consider that no events occur in the first 30 frames, from which we derive the baseline distribution's parameters, $\mu = -0.333$, $\sigma_1^2 = 35.54$. We estimate the parameters for P_1 with a window of 50 frames. As an illustrative example, we focus on the distribution of the illumination variations of one pixel $\bar{r}_0$ inside the activity area, with $D^1(\bar{r}_0)$ shown in Fig. 2(b). The log-likelihood ratio's values (Fig. 2(c)) indeed change around the frames where pixel $\bar{r}_0$ starts and stops moving. The maxima of its gradient, shown in Fig. 2(d) indicate that a change occurs

[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

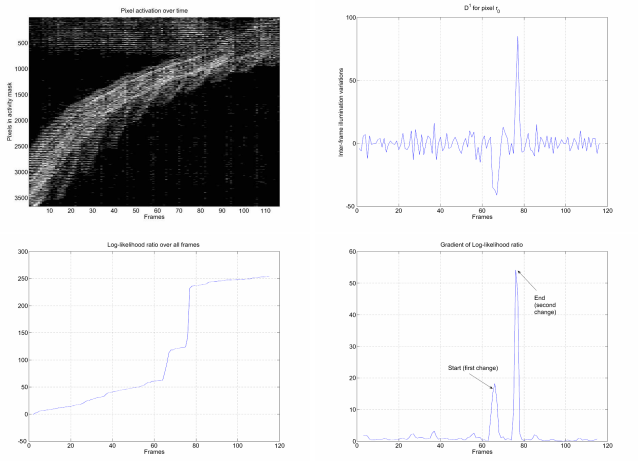

Fig. 2. Walking. (a) Luminance variations in activity mask. (b) Activity mask. (c) Log-likelihood ratio of active pixel's luminance variations. (d) Gradient of log-likelihood ratio.

(it becomes active) at frame 66 and a new change occurs at frame 76. A similar procedure is repeated for all frame pixels, for the extraction of frames at which pixels undergo change. The accuracy of these results is verified by examining the video sequences and finding at which frames changes take place. The observed changes indeed coincide with those found by our algorithm (with small discrepancies of one or two frames), and led to a high detection rate, equal to 94%.

Meet-Walk Sequence. This video showed two people walking towards each other, meeting and then walking away together (Fig. 3). The activity area is interesting, as it contains their trajectories, but also reveals the motion of some other people walking in the back of the room (top left of activity area in Fig. 3(d)).

We consider that no events occur in the first 30 frames, and derive $\mu = -0.02$, $\sigma_1^2 = 2.13$ for P_0. A window of 50 frames is then used to estimate the parameters for P_1 . The variations of all active pixels' illumination is shown in Fig. 4(a). The illumination variations $D^1(\bar{r}_0)$ of a pixel $\bar{r}_0$ in the activity area are shown in Fig. 4(b). The log-likelihood ratio in Fig. 4(c) indeed has values that increase around the frame where change occurs. The frame of change is found from the gradient (Fig. 4(d)), which shows that a change occurs at frame 164. As before, the ground truth was obtained by visually observing the videos. Our algorithm led to a 92% accuracy in the detection of the changes in these activity areas, so it can indeed be considered a reliable measure of change.

5. CONCLUSIONS

A statistical method is presented for the spatiotemporal localization of activity in video. Binary masks of active pix-

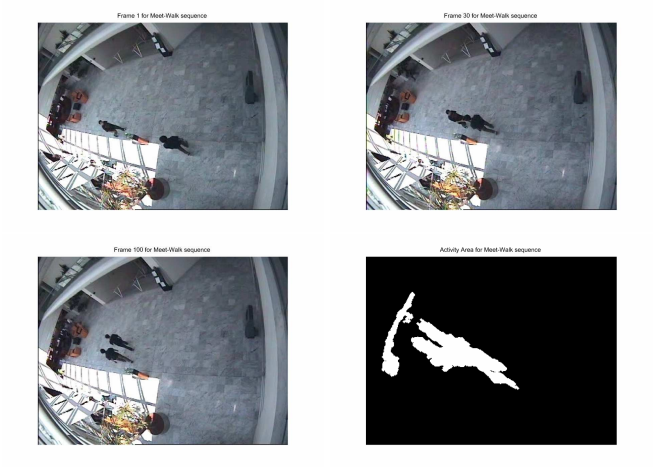

Fig. 3. Meet, Walk. (a) Frame 1. (b) Frame 30, meet. (c) Frame 100, walk away. (d) Activity mask.

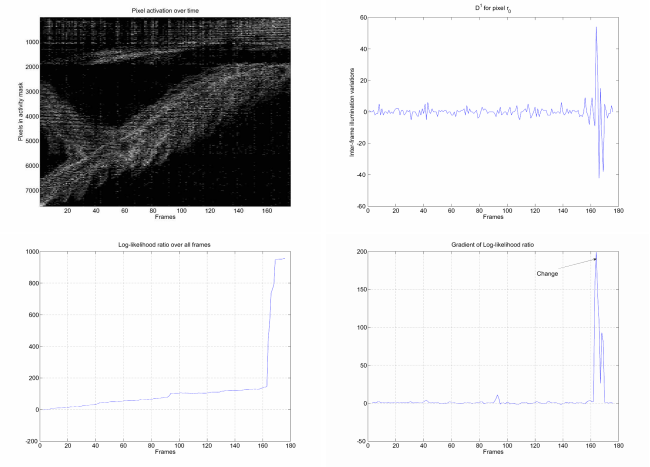

Fig. 4. Meet, Walk. (a) Luminance variations in activity mask. (b) Activity mask. (c) Log-likelihood ratio for active pixel's luminance variations. (d) Gradient of log-likelihood ratio.

els in the video are extracted from the higher order statistics of luminance variations over all frames. The activity masks are then used to use only the active pixels to detect times of change, providing lower computational cost and higher system reliability. The statistics of each active pixel's inter-frame illumination variations are processed via sequential likelihood ratio testing to detect changes in it. The changes correspond to the beginning or ending of events. Experiments compared the activity masks with human observations of the location of activity in the video and showed that they are accurate. Finally, the ground truth for times of change was obtained by observing the video. These values were compared with the times of change extracted from sequential likelihood ratio processing, and were found to lead to high probabilities of detection. Future work involves the use of empirical models for the data distribution in the likelihood ratios, rather than the Gaussian approximation, as well as more detailed examination of the semantics of the changes detected in videos.

6. REFERENCES

[1] G.L. Foresti, C. Micheloni, L. Snidaro, P. Remagnino, and T. Ellis, "Active video-based surveillance system: the low-level image and video processing techniques needed for implementation," *Signal Processing Magazine, IEEE*, vol. 22, no. 2, pp. 25 – 37, March 2005.

[2] W.-H. Lin and A. G. Hauptmann, "News video classification using svm-based multimodal classifiers and combination strategies," in *Proc. ACM Multimedia*, Juanles-Pins, France, 2002.

[3] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257–267, March 2001.

[4] J.L. Barron and R. Eagleson, "Recursive estimation of time-varying motion and structure parameters," *Pattern Recognition*, vol. 29, no. 5, pp. 797–818, Dec. 1996.

[5] B.K.P. Horn and B.G. Schunck, "Determining optical flow," *Artificial Intelligence*, vol. 17, pp. 185–203, 1981.

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

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

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

[9] A.K. Nandi, "Robust estimation of third-order cumulants in applications ofhigher-order statistics," *Radar and Signal Processing, IEE Proceedings*, vol. 140, no. 6, pp. 380–389, Dec. 1993.

[10] 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*, Sept. 1998, pp. 411 – 414.

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