Robust Temporal Activity Templates Using Higher Order Statistics

Alexia Briassouli and Ioannis Kompatsiaris, *Member, IEEE*

Abstract

A robust, theoretically-founded approach for the extraction of temporal templates corresponding to areas of motion in video, is presented. Higher order statistics (kurtosis) are employed to extract Activity Areas, i.e. binary masks indicating which pixels in a video are active. The application of the kurtosis on illumination changes modeled as Gaussians and Mixture of Gaussians is shown to be sensitive to outliers for both models, thus correctly localizing active pixels. Activity Areas are compared to existing, difference-based temporal templates, known as Motion Energy Images, and the robustness of both categories of temporal templates to additive noise is analyzed theoretically. Experiments with numerous real videos with additive noise, both indoors and outdoors, are conducted to compare the robustness of the Activity Areas and Motion Energy Images, and their temporal extensions, the Activity History Areas and Motion History Images. As expected from the theoretical analysis, the kurtosis-based Activity Areas prove to be more robust than the difference-based templates. Challenging videos containing occlusions, varying backgrounds and shadows are also examined, and it is shown that the proposed approach outperforms the difference-based method for these cases as well, consistently providing reliable localization of activity under a wide range of difficult circumstances. The proposed approach provides good results at a very low computational cost, and without requiring prior knowledge about the scene, nor training of any kind.

EDICS Category: ARS-IIU

Manuscript received October 10, 2008; revised May 18, 2009. This work was supported by the European Communitys Seventh Framework Programme FP7/2007-2013 under grant agreement FP7-214306 - JUMAS, from FP6 under contract number 045547-VidiVideo. A. Briassouli and I. Kompatsiaris are with the Informatics and Telematics Institute, CERTH-ITI, Thessaloniki, Greece. (e-mail: abria@iti.gr, ikom@iti.gr)

I. INTRODUCTION

The analysis of video data has received significant research attention in numerous applications, such as video surveillance, activity recognition, event detection. A fundamental initial step in video processing is the analysis of the motion present in it, detected as changes in its frames' illumination. Numerous methods have been developed for the estimation of motion in video, as well as its further processing for the extraction of practically useful results, such as the characterization of human activity, or the detection of anomalous events. For this purpose, the detection of regions of activity in a video, and the separation of foreground objects from the background, are of fundamental importance.

The analysis of motion in video often extracts binary masks of pixels that are active over several frames, creating a foreground mask for them, which can be considered as an extension of background removal [1]. The pixels that are active in the sequence being examined form a binary mask, which will be referred hitherto as a temporal template, in the general case. In the literature, the temporal templates that have been examined can be either a binary mask or a gray-scale image, which indicates the temporal history of pixel motions, having higher values at the pixels that have been most recently activated [1], [2]. This work is extended in [3], where the shape of the temporal template is examined in the same spatial location in each image, and mapped onto an eigenspace.

In practice, temporal templates find applications in tasks such as human action recognition or classification, because the regions of active pixels often have a shape that is characteristic of the activity taking place. In both [1], [3], good recognition results are obtained from the use of these templates, but the issue of the system's robustness to noise is not examined, and only simple, noise-free indoors sequences, with nearly no background, are examined. In [4], [5], video is treated as a spatiotemporal volume and a different kind of temporal templates (space-time features) are extracted and used for recognition purposes. This approach gives good results, but is limited to human action recognition, as it extracts features that are characteristic of the human body and its motions. There exist many other techniques for the recognition of activities, or events, such as HMM-based methods [6], [7], which involve training for the modeling of the event or activity. Other approaches are based on the analysis of the motion statistics [8], [9], but are better suited for the classification of motions with varied statistical distributions of motion vectors (e.g. a news video versus a traffic scene), or the analysis of trajectories, rather than the recognition of specific activities, that lead to temporal templates with a characteristic shape.

In order to extract accurate temporal templates, it is important to separate active from static pixels with precision at each frame, i.e. to perform accurate background subtraction. In the case of a moving

camera, camera motion compensation needs to take place before the foreground-background separation. In the literature, there has been very extensive work on this problem, as it is fundamental in many applications. A very popular approach involves the modeling of the background as a Mixture of Gaussians [10], [11], and uses an adaptive scheme to account for varying scene illumination and small background motions. However, these techniques require the correct choice of parameters for the model, and are often computationally costly. Most importantly, they often suffer from slow adaptation to scene illumination changes and can incorporate foreground objects into the background. An approach for the detection of activity in video that is robust to illumination changes has been proposed in [12]. In that work, the scene illumination changes are estimated and can thusly be accounted for. However, the estimation of scene illumination changes is made possible by using prior knowledge of the locations of these changes. In practice, this knowledge is usually not known, and can even be part of the desired output of the system. Another background extraction method that is robust to sudden illumination changes is presented in [13]. Similarly to [12], this method requires prior knowledge, as it uses a stored set of background images for the modeling of the illumination changes at the current frame. Additionally, the work in [13] also involves the manual extraction of foreground areas during training for the modeling of the background and foreground distributions.

A different set of approaches to the problem of background estimation is based on non-parametric modeling of the background-foreground data using kernels [14], [15]. These approaches lead to good results without requiring prior knowledge of the scene, and can separate the foreground from the background for challenging videos containing moving backgrounds (e.g. waves) and moving objects at the same time. However, accurate non-parametric modeling is very computationally expensive in terms both of speed and memory requirements.

Small background variations and even small camera motions have also been handled by using local flow measurements, which are separated from the dominant scene motion. In [16], the (registered) video frames are temporally integrated, leading to the smoothing out of small camera motions and background variations and leading to the localization of the foreground objects. In [17], camera motion is first compensated for, and the resulting residual motion is used to obtain local motion estimates, which correspond to the actual foreground motion in the scene. These approaches have not been examined for their robustness to global scene illumination changes. Additionally, by their nature, local techniques can become sensitive to spatially or temporally occurring local noise, such as local occlusions, illumination changes, shadows.

Robustness to local sources of noise can be achieved, in general, by using global methods, i.e. methods that simultaneously employ many - or all - frame pixels and video frames. The use of this additional

data increases the robustness of the system, as the detrimental effect of locally erroneous values, such as those caused by small measurement errors, camera instabilities, local occlusions, is compensated for by the incorporation of correct values from the rest of the spatio-temporal video segment being used (which can even be the entire video). In this work, we present a global, computationally efficient method for the derivation of activity areas based on higher order statistical processing of inter-frame illumination changes. The measurement noise in static pixels can be approximated by the Gaussian distribution, as is often the case in the literature [18], [19], [20]. The kurtosis of Gaussian noise is zero, so the kurtosis of illumination changes occurring on background pixels should have low values, while its value for active pixels is expected to be much higher. The robustness of the proposed method to additive noise is analyzed theoretically, and the effect of additive noise is shown to be small. The application of the kurtosis to data modeled as Mixtures of Gaussians is also analyzed, as these distributions are often used to represent background pixels [10]. It is shown that even in that case (where the background is not modeled by a single Gaussian), the higher order statistical measure obtains low values in static pixels. The proposed approach is compared with the often used difference-based temporal templates, namely the Motion Energy Images: theoretical analysis of that method's robustness to additive noise shows that it is not inherently unaffected by noise, unlike the proposed approach.

The use of higher order statistics is motivated by the fact that they have been proven to be sensitive to outliers, and are thus expected to distinguish reliably between illumination changes induced by measurement noise and those caused by actual motion, even when the background illumination variations do not follow a strictly Gaussian distribution. In the literature, [21], [22], fourth order statistics (the kurtosis) have been used for motion estimation and noise removal based on the minimization of fourth order cumulants of the estimation errors. It is shown that fourth order cumulants lead to more accurate results than the minimization of the mean squared error, even when the noise is colored Gaussian or non-Gaussian. This is due to the fact that the kurtosis remains asymptotically unaffected by zero-mean additive noise of any distribution [23]. Thus, in our case it is also expected to accurately distinguish between the random noise distribution and outlier values caused by motion.

Extensive experimentation with a large variety of indoors and outdoors videos for increasing amounts of additive noise demonstrates that higher order statistics provide much more stable temporal templates that the difference-based approach. Post-processing is applied to the results of both methods for the case where shadows are present, leading to their effective elimination in both cases. The effect of using temporal windows of varying size for the extraction of temporal templates is also examined, for both the kurtosis and difference-based methods. It is determined that the effect of varying window size is very small, and

that the active pixels are still extracted with accuracy. Finally, experiments with challenging videos that contain background motions (e.g. moving textures like trees moving in the wind, waves), global scene illumination changes and shadows demonstrate the effectiveness of this approach in difficult, realistic circumstances.

It should be emphasized that, although the kurtosis-based processing is robust to noise, moving background textures and varying illumination, it has a low computational complexity, very similar to that of the difference-based approach. Additionally, the proposed approach does not make use of any prior knowledge, but only processes the video frames themselves. Its robustness to the amount of frames used for its computation (varying temporal window size) allows its computation for selected video subsequences, which may contain characteristic segments of the activity taking place, instead of for the entire video. Finally, its global nature increases its insensitivity to local sources of noise.

II. Kurtosis-based Activity Area

The region where activities occur in a video often has a characteristic shape which can be used to classify, for example, the type of motion taking place. This region can be derived as a binary mask, with value equal to 1 at the pixels where motion has occurred, and 0 otherwise. Such a mask can be created based on the value of the illumination changes in a video that has frame luminance values $I(x, y, t)$. It should be emphasized that this kind of a mask is meaningful only when the camera is static, or camera motion has been compensated for. Thus, in the most general case, it should be derived from the residual motion vectors (the motion measurements remaining after camera motion compensation), as in [17]. In many cases where the camera is not moving, inter-frame differencing gives a sufficiently accurate indication of motion-induced illumination changes between pairs of frames, with very few false alarms. For example, indoors videos (e.g. in surveillance) are often filmed with static cameras, whose quality has improved significantly in recent years, so frame differences correspond to actual motions. For more challenging cases where there is slight camera motion, panning, or an increased amount of camera noise, the illumination changes between pairs of frames can be approximated by flow estimates [24], [25] after motion compensation. For simplicity, we refer to either inter-frame illumination differences or optical flow measurements as "illumination changes", denoted as $dI(x, y, t)$:

$$dI(x, y, t) = I(x, y, t) - I(x, y, t - 1). \tag{1}$$

In practice, real videos contain slight illumination changes, introduced by camera instability, non-constant camera exposure or other sources of measurement noise [26], so when illumination differences are accumulated over a sequence of frames, some correspond to actual motion, and some to noise.

The measurement noise, which affects all frame pixels in an additive manner, is often modeled in the literature as a Gaussian distribution [19], [18]. This is the simplest model of the background pixels' distribution, with more complex ones having been proposed in the literature, like the well known mixture of Gaussians [10]. Despite its simplicity, the Gaussian assumption can lead to reliable and robust results, particularly in the case examined in this work, since the statistic being used, namely the kurtosis, is robust to deviations of the data from a strictly Gaussian model [21], [27]. It should also be noted that the amount of measurement noise samples available is usually high, making it possible to model them by the Gaussian Distribution, based on the Central Limit theorem [28], [29]. For example, a video with N frames of size $N_1 \times N_2$ provides $N_1 \times N_2 \times (N-1)$ samples: in our experiments, videos, with 100 frames of size 120×160, provide $1.900.800$ data samples.

The illumination changes caused in a pixel by actual motion and not solely by measurement noise introduce an outlier value to the otherwise random measurement noise [30]. The illumination changes then deviate significantly from the Gaussian model, since object motion is very different from random illumination changes and measurement noise. If the illumination changes are represented by $dI(x,y,t)$ at pixel (x,y) and frame t, they can be mapped to the following two hypotheses:

$$\begin{aligned} H_0 : dI^0(x,y,t) &= z(x,y,t) \\ H_1 : dI^1(x,y,t) &= u(x,y,t) + z(x,y,t). \end{aligned} \tag{2}$$

where $dI^0(x,y,t)$ in Hypothesis H_0 represents the illumination change at pixel (x,y), in frame t introduced by measurement noise $z(x,y,t)$, and H_1 corresponds to the case where there is also motion at pixel (x,y), represented by $u(x,y,t)$. Under the Gaussian approximation for the measurement noise $z(x,y,t)$, we can detect which velocity estimates correspond to a pixel that is actually moving by examining the non-Gaussianity of the accumulated velocity estimates [31]. The non-Gaussianity of a random variable y can be tested using higher order statistics, i.e. the $k-$order moments of a random variable y, defined as:

$$m_k = E[y^k] = \int y^k f_y(y) dy, \tag{3}$$

where $f_y(y)$ is the probability density function of the random variable y. These moments are referred to as the "raw moments" of y because they are not centered around its mean, but they coincide with the raw moments if the mean of y is simply subtracted from its values. The kurtosis of y is defined as:

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

The fourth moment of a Gaussian random variable is $m_4 = E[y^4] = 3(E[y^2])^2$, so its kurtosis is equal to zero. Therefore, the kurtosis is a natural choice for finding which pixels undergo illumination changes that do not follow a Gaussian distribution, i.e. which pixels are active. The kurtosis has also been shown [32], [33], [27] to be a robust detector of outliers in both Gaussian and non-Gaussian data [34]. Thus, it is very suitable for detecting active pixels in video, as these introduce large illumination changes in the data, that are outliers compared to the values of the measurement noise.

A. Kurtosis in the presence of noise

In order to examine the sensitivity of the kurtosis to additive noise, we consider zero-mean additive noise v (w.l.o.g., since the mean can be subtracted from our data) and a Gaussian random variable y:

$$K_{y+v} = E[(y+v)^4] - 3(E[(y+v)^2])^2, \tag{5}$$

with

$$\begin{aligned} E[(y+v)^4] &= E[(y^2+v^2+2yv)^2] = E[y^4] + E[v^4] + 6E[y^2v^2] + 4E[yv(y^2+v^2)] \\ &= E[y^4] + E[v^4] + 6E[y^2v^2] + 4(E[y^3v] + E[yv^3]) = E[y^4] + E[v^4] + h.o.t, \end{aligned} \tag{6}$$

with higher order terms $h.o.t = 6E[y^2v^2] + 4(E[y^3v] + E[yv^3])$. Also:

$$E[(y+v)^2] = E[y^2] + E[v^2] + 2E[yv] = E[y^2] + E[v^2] + 2E[y]E[v] = \sigma_y^2 + \sigma_v^2, \tag{7}$$

where we assume that the data y and the additive noise v are independent, i.e. $E[yv] = E[y]E[v]$. This assumption can be made because the additive noise under consideration is externally originating, so it is independent of the imaging process (unlike measurement noise). Since $y,\ v$ are zero-mean, $E[yv] = E[y]E[v] = 0$, and $E[y^2] = \sigma_y^2,\ E[v^2] = \sigma_v^2$, where $\sigma_y^2,\ \sigma_v^2$ represent the variance of the data and the additive noise respectively. Then Eq. (5) becomes:

$$K_{y+v} = E[y^4] + E[v^4] + h.o.t. - 3(E[y^2])^2 - 3(E[v^2])^2 - 6E[y^2]E[v^2] = K_y + K_v + H.o.t., \tag{8}$$

where $K_y = E[y^4] - 3(E[y^2])^2,\ K_v = E[v^4] - 3(E[v^2])^2$ are the kurtosis of the data and the Gaussian noise, respectively and $H.o.t = h.o.t. - 6E[y^2]E[v^2]$. The kurtosis of y is equal to zero, since y is a Gaussian random variable, so the kurtosis of the noisy data of Eq. (9) becomes:

$$K_{y+v} = K_y + K_v + H.o.t = K_v + H.o.t. \tag{9}$$

In practice, the higher order terms are often negligible, as the quantities in them obtain very low values: the additive noise values v are usually small (less than one), and when raised to the second or third

power, become even smaller, so that $E[y^2v^2] \simeq 0,\ E[y^3v] \simeq 0,\ E[yv^3] \simeq 0$, making $H.o.t \simeq 0$ and $K_{y+v} = K_v$. Then, the kurtosis of Gaussian data in the presence of non-Gaussian additive noise is equal to the kurtosis of this noise v. Naturally, if the additive noise is Gaussian, and under the assumption that y and v are independent, the kurtosis of $y+v$ becomes zero again, since $K_v = 0$. This is expected, since the sum of independently distributed Gaussian random variables ($y + v$) is also Gaussian [28], [35].

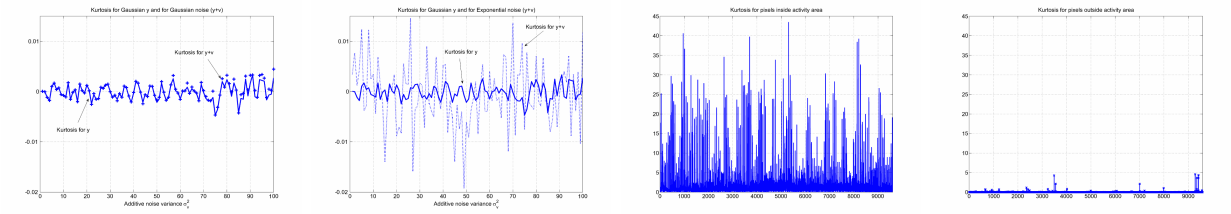
It should be noted that, although in most practical cases the higher order terms (H.o.t.) can be set to zero, leading to $K_{y+v} = K_v$, they cannot be considered negligible under very high additive noise, where their influence becomes visible in the results. This can be seen in the experiments for high amounts of additive noise. However, videos with such amounts of noise are extremely degraded and not useful in practice. Consequently, for most realistic situations, the above approximation can be considered valid.

B. Monte-Carlo Simulations of kurtosis in the presence of Gaussian and non-Gaussian noise

In order to experimentally demonstrate how the kurtosis of a Gaussian and noise-corrupted Gaussian random variable deviates from zero as a function of the additive noise variance σ_v^2, we use Monte-Carlo simulations. We run 1000 simulations where we generate 10^6 samples of a random variable y following a zero-mean Gaussian distribution with variance equal to one, and estimate its kurtosis. The first point of Fig. 1(a) shows the values of its kurtosis for zero noise: it is evident that the kurtosis is near zero. Zero-mean Gaussian noise v, with variance that increases from 0 to 1 by 0.01 increments is then added to the original random variables y, and the kurtosis values of $y + v$ are also plotted in Fig. 1(a): as expected, the values of the kurtosis remain low, between ± 0.004.

In order to compare this with the kurtosis behavior under non-Gaussian noise, the same experiments are performed, but this time with the addition of Exponentially [30], [36] distributed noise v, that has the same energy as the previously used Gaussian noise, for fairness of comparison. In Fig. 1(b) we see that under increasing non-Gaussian noise, the kurtosis obtains values between ± 0.02, i.e. they increase by an order of magnitude. This is actually an advantage in our case, since it implies that the kurtosis will deviate from zero (or low values) with the addition of non-Gaussian object velocities to the measurement noise under H_1 (Eq. (2)).

We also verify that the kurtosis is a reliable measure of pixel motion for a real video sequence, where we examine the behavior of the kurtosis in moving and non-moving frame pixels. We use a video of a person running (Sec. VI-C) and prior knowledge about which pixels are moving and not moving, since this experiment is designed to simply demonstrate the behavior of the kurtosis in a real example. We then estimate the kurtosis of all *active* pixels over time (i.e. over all video frames) and the kurtosis of the static



(a) (b) (c) (d)

Fig. 1. Kurtosis of random variable y and noisy random variable $y+v$, where v is: (a) Gaussian. (b) Exponential. The kurtosis of the Gaussian random variable is nearly zero and remains close to zero for Gaussian noise. Under non-Gaussian noise, the values of the kurtosis increase significantly. Kurtosis of the pixel illumination changes averaged over time for: (c) pixels inside the activity area, (d) pixels in the rest of the video frames. The static pixels have a non-zero kurtosis, but with much lower values that that of the active pixels.

(or near-static) ones, also over time. It should be noted that the expectations $E[\cdot]$ in Eq. (4) and Eq. (5) are approximated by averaging over all moving and non-moving pixels, respectively. This is because we make the assumption that the random variables being measured correspond to an ergodic process (due to lack of additional knowledge), so their expectations are approximated by arithmetic means [36], [37]. In practice this approximation proves adequate, as it leads to reliable estimates of the data statistics. Fig. 1(c), (d) shows that, indeed, the kurtosis values deviate significantly from zero in the moving pixels. We have plotted the values of the kurtosis on the same scale to emphasize that they increase by orders of magnitude in the active pixels, even if it is not guaranteed that the additive noise strictly follows a Gaussian distribution. Indeed, the average of the kurtosis values of the moving pixels is equal to 2.1941, whereas the static pixels have a mean kurtosis equal to 0.0058. Thus, there is a significant deviation between the kurtosis of static and active pixels, which indicates that in practice, it will be a reliable measure of pixel activity, even for noise that is not strictly Gaussian.

C. Kurtosis-based Activity Areas from Illumination Changes

The results of Sec. II-A, II-B motivate us to use the kurtosis for the spatial localization of activity. In Appendix I, the kurtosis of the illumination changes $dI(x,y,t)$ is shown to be equal to:

$$K_{dI} = 2K_I + K_c, \tag{10}$$

where m_k are the k order statistics of each frame's illumination values, defined in Eq. (3), considered to be the same for all video frames, and K_c are cross-correlation terms. If the additive noise follows a

Gaussian distribution $\mathcal{N}(0, \sigma_n^2)$, the noise differences $dn(x,y,t) = n(x,y,t) - n(x,y,t-1)$ also follow a Gaussian distribution [28] $\mathcal{N}(0, 2\sigma_n^2)$, so their kurtosis is zero, i.e. $K_{dn} = 0$.

D. Effect of noise on Kurtosis-based Activity Areas from Illumination Changes

In order to examine the robustness of the kurtosis-based activity areas to occlusions, camera jitter, illumination changes, we model these effects as additive noise, and then estimate the kurtosis of noisy inter-frame illumination differences, $\delta(x,y,t) = (I(x,y,t) + n(x,y,t)) - (I(x,y,t-1) + n(x,y,t-1))$. For notational simplicity, we do not include the indices $(x,y,t),\ (x,y,t-1)$ in the equations that follow, but use: $dI = I(x,y,t) - I(x,y,t-1)$ and $dn = n(x,y,t) - n(x,y,t-1)$ as in Sec. II-C. Then:

$$\begin{aligned} \delta^4 &= (dI + dn)^4 = (dI^2 + dn^2 + 2dI \cdot dn)^2 = dI^4 + dn^4 + 6dI^2dn^2 + 4dI^3dn + 4dIdn^3 \Rightarrow \\ E[\delta^4] &= E[dI^4] + E[dn^4] + 6E[dI^2dn^2] + 4E[dI^3dn] + 4E[dIdn^3]. \end{aligned} \tag{11}$$

Also:

$$(E[\delta^2])^2 = (E[dI^2] + E[dn^2] + 2E[dIdn])^2 = (E[dI^2])^2 + (E[dn^2])^2 + 2E[dI^2]E[dn^2], \tag{12}$$

where we have considered that dI and dn are assumed to be independent, since the additive noise n is externally generated (i.e. it is unrelated to the measurement noise and the imaging process in general). Then $E[dIdn] = E[dI]E[dn] = 0$ since $E[dn] = 0$. The kurtosis is $K_\delta = E[\delta^4] - 3E^2[\delta^2]$, so Eqs. (11), (12) lead to:

$$\begin{aligned} K_\delta &= E[dI^4] + E[dn^4] + 6E[dI^2dn^2] + 4E[dI^3dn] + 4E[dIdn^3] - 3(E[dI^2])^2 - 3(E[dn^2])^2 - 6E[dI^2]E[dn^2] \\ &= K_{dI} + K_{dn} + 6(E[dI^2dn^2] - E[dI^2]E[dn^2]) + 4E[dI^3dn] + 4E[dIdn^3] \\ &. = K_{dI} + K_{dn} + 6(E[dI^2dn^2] - E[dI^2]E[dn^2]), \end{aligned} \tag{13}$$

since $E[dn] = E[dI] = 0$. The term $6(E[dI^2dn^2] - E[dI^2]E[dn^2])$ is equal to zero if the higher powers of dI, dn are also considered to be independent. This assumption is reasonable due to the fact that dI and dn originate from completely different and independent sources. Then, we have:

$$K_\delta = K_{dI} + K_{dn}, \tag{14}$$

Eq. (14) has a central role in demonstrating the robustness of the kurtosis for the extraction of the activity area. If the additive noise $n(x,y,t)$ follows a Gaussian distribution, $K_{dn} = 0$ and:

$$K_\delta = K_{dI}. \tag{15}$$

In other words, the kurtosis of the illumination changes remains unaffected by additive Gaussian noise! Additionally, even when the additive noise cannot be modeled by a strictly Gaussian distribution, the kurtosis of the illumination changes remains sensitive to outliers [21], [22]. In [32], for example, it is shown that the kurtosis can still reliably detect outliers, even when they are embedded in generalized Gaussian noise. In the section that follows, we analyze mathematically how the kurtosis behaves under Mixture of Gaussian (MoG) modeling of the data. This allows us to see how the deviation of the background from a (potentially) more accurate model than the single Gaussian will affect the accuracy of finding active pixels with the kurtosis.

III. Kurtosis-based activity areas for Mixture of Gaussian modeling

A very commonly used model for the background is based on mixtures of Gaussian (MoG) distributions. Numerous improvements have been developed for the modeling of background using mixtures of Gaussians, e.g. adaptive modeling as in [10] and [11] to account for small background motions like leaves fluttering, small illumination changes, shadows. Since this model is so widely used, we examine how the kurtosis-based activity areas are affected when the data is modeled by a MoG. The pdf for a random variable X modeled by a MoG is given by the convex sum of N Gaussian distributions f_i with mean μ_i and variance σ_i^2, i.e. $\mathcal{N}(0, \sigma_i^2)$: $f_X(x) = \sum_{i=1}^{N} w_i f_i(x)$, where the weights $w_i \in (0, 1)$ are such that $\sum_{i=1}^{N} w_i = 1$. The convexity of the sum leads to the higher order moments of X:

$$E[X^k] = \sum_{i=1}^{N} w_i E_{fi}[X^k], \tag{16}$$

where $E_{fi}[X^k]$ are the k^{th} order moments of the individual Gaussian distributions f_i. The kurtosis for data modeled by a MoG distribution is estimated analytically in Appendix B, and found to be:

$$K_X = 3\Bigg(\sum_{i=1}^{N} w_i(1-w_i)E_{fi}^2[X^2]) - \sum_{i\neq j} w_i w_j E_{fi}[X^2]E_{fj}[X^2]\Bigg). \tag{17}$$

Since the static pixels' illumination is modeled by a MoG, the variable X is equal to the pixel illumination I, i.e. $X = I$. However, we apply the kurtosis measure to illumination changes, i.e. dI. Their kurtosis for MoG background modeling, is then found from Eqs. (10), (17) to be:

$$K_{dI} = 6\Bigg(\sum_{i=1}^{N} w_i(1-w_i)E_{fi}^2[I^2]) - \sum_{i\neq j} w_i w_j E_{fi}[I^2]E_{fj}[I^2]\Bigg) + K_c. \tag{18}$$

The resulting value of the kurtosis contains the products and higher powers of the weights w_i, where $0 < w_i < 1$, which will attenuate its value. Indeed, our experiments show that the kurtosis of the background pixels, in many different videos, filmed both indoors and outdoors, obtains much lower values

than the active pixels. This is expected, as the kurtosis has been proven to be asymptotically unaffected by zero-mean additive noise, even when it does not follow a strictly Gaussian distribution [38]. Thus, if MoG data (e.g. the background values) is regarded as Gaussian, contaminated by unknown noise, the kurtosis will still converge to zero asymptotically (and to low values in practical applications). Indeed, in [21] higher order statistics, and specifically fourth order cumulants are actually used to successfully de-noise image sequences, without modeling the noise as following a strictly Gaussian distribution.

IV. Difference-Based Motion Energy Image

In this section, we present the "temporal templates" proposed in [1] for the characterization of pixels as moving or static, similarly to the activity areas of Sec. II. These templates are created based on the value of the illumination changes in a video. In [1], illumination changes $d(x, y, t)$ are first binarized, leading to the values $D(x, y, t)$, which are equal to 1 when pixel (x, y) has undergone motion between frames t and $t - 1$, and zero otherwise. The illumination changes are binarized, using a threshold η (Sec. IV-A), as follows:

$$D(x, y, t) = \begin{cases} 1, & if\ d(x, y, t) \geq \eta; \\ 0, & else. \end{cases} \tag{19}$$

The binary MEI is then defined as follows:

$$E_\tau(x, y, t) = \bigcup_{i=0}^{\tau-1} D(x, y, t - i), \tag{20}$$

where $\bigcup_{i=0}^{\tau-1}$ denotes the union of the binarized illumination changes from frames 0 to $\tau - 1$ and the parameter τ determines the duration of the activity being examined. Essentially, Eq. (20) shows that the MEI is a concatenation of the active pixels throughout the τ frames under examination, with value 1 in the active pixels and 0 otherwise.

A. Threshold Selection for MEI and Activity Area

In order to create the MEI from the illumination differences, the threshold η of Eq. (19) needs to be determined in a generally applicable, non-ad-hoc manner. In this section, a threshold that is shown to give excellent experimental results is presented. It is chosen based on Chebyshev's inequality, which holds for a random variable x with finite mean μ, variance σ^2, that follows any distribution:

$$P(|x - \mu| \geq k\sigma) \leq \frac{1}{k^2}. \tag{21}$$

For $k = 3$, this shows that at least 89% of data belongs within three standard deviations from its mean. Thus, the threshold for binarizing the illumination changes dI and estimating the MEI is given by:

$$\eta = \mu_{d_I} + 3\sigma_{d_I}, \tag{22}$$

where μ_{d_I} is the mean of the illumination changes and σ_{d_I} the standard deviation of their distribution. In order to estimate these quantities, we need to make the following assumptions:

1) Frame statistics do not change significantly, so k^{th} order moments for all frames are equal to m_k.
2) Illumination values of successive frames are statistically independent of each other.

The first assumption is realistic when there are no large global illumination changes between frames and when the motion between them is not very high, but will not hold when there are large occlusions and dis-occlusions, if different objects enter the scene or there are large changes in global illumination. In most realistic cases, like the indoors and outdoors videos examined in our experiments, deviations from this assumption do not introduce significant degradations in the results (see Sec. VI). Additionally, the proposed method uses many (or all) video frames, so that errors in a subset of the data, caused for example by a sudden illumination change, are often compensated for by the presence of many correct samples. The second assumption is not as realistic, since pixel values of successive video frames are not independent. Nevertheless, similar assumptions are often made in practice to enable the theoretical performance analysis, here of the proposed temporal templates. Most importantly, the errors introduced in practice by this assumption are practically negligible, as shown by our experiments. Then:

$$E[d_I(x, y, t)] = \mu_{d_I} = E[I(x, y, t) - I(x, y, t - 1)] = 0, \tag{23}$$

where we take into account the first assumption, that the mean m_1 of the illumination values in frames t and $t - 1$ is the same. $E[\cdot]$ is estimated over all frame pixels (from our assumption of spatial-ergodicity, due to the lack of additional knowledge about the data). Under the second assumption, the illumination values of successive frames are independent, so the variance of $dI = I(x, y, t) - I(x, y, t - 1)$ is:

$$\sigma^2_{d_I}(x, y, t) = 2(m_2 - m_1^2) = 2\sigma_I^2. \tag{24}$$

Using Eq (23) and (24), the threshold of Eq. (22) for obtaining the MEIs becomes:

$$\eta = 3\sqrt{2}\sigma_I, \tag{25}$$

where σ_I^2 is the variance of the illumination values in each frame (estimated over the entire frame's pixels). It should be noted that the kurtosis values, estimated over each pixel of sequence of frames, lead to the binary activity areas by thresholding with Eq. (22), using the mean and variance of the kurtosis values instead of the mean and variance of the illumination changes.

B. MEI in the presence of noise

We examine how the MEI is affected by additive noise $n(x, y, t)$ (at pixel (x, y), frame t), assumed w.l.o.g. to be zero-mean and with variance σ_n^2. Using the notation of Sec. II-D, the noisy data is $\delta = dI + dn$, where dI are inter-frame illumination changes and dn the inter-frame variations of the additive noise. As before, we make the assumption that inter-frame illumination changes dI are independent from additive noise changes dn, since they originate from different sources. Then, $E[dI] = E[dn] = 0$, so $E[\delta] = 0$ and $\sigma_\delta^2 = \sigma_{dI}^2 + \sigma_{dn}^2 = 2(\sigma_I^2 + \sigma_n^2)$, where the expected values are spatial averages. Additive noise increases the data variance and, since the MEI is formed based on these second order statistics, it is expected to be affected by it. In practice Eq. (22) becomes

$$\eta_n = 3\sqrt{2(\sigma_I^2 + \sigma_n^2)}. \tag{26}$$

Although the threshold increases in the presence of noise, the MEI is nonetheless adversely affected by it, as verified by our extensive experimental results (Sec. VI). Additive noise is always present in real videos, indoors and outdoors: for instance, there may be inter-frame illumination changes, camera measurement noise (thermal noise, photon noise), small background movements (e.g. tree leaves in the wind) etc. The MEIs mistake this noise for motion, so it appears in the resulting binary template images. This introduces errors when extracting temporal templates for classification and recognition of the activities taking place.

V. Motion History Image, Activity History Area

In [1], the Motion History Image (MHI) is also used to complement the information of the Motion Energy Image (MEI), for describing the temporal evolution of the motion, i.e. which parts of the motion template (MEI) are activated sooner and which later. The MHI's are extracted from the binarized frame illumination differences [1] as follows:

$$H_\tau(x, y, t) = \begin{cases} \tau, & if\ D(x, y, t) = 1; \\ max(0, H_\tau(x, y, t-1) - 1), & else. \end{cases} \tag{27}$$

This results in grayscale masks with higher values indicating the regions that were active more recently. We define the Activity History Area (AHA) in a similar manner to the MHI, but the active pixels that are included in it are extracted using binarized kurtosis values instead of binarized inter-frame illumination differences. For a video with frames of size $N_1 \times N_2$, the kurtosis of each pixel's illumination changes over t frames is estimated from Eq. (10). The expected values are estimated for each pixel over time, leading to a $N_1 \times N_2$ image of the kurtosis values. These are then binarized (as in the end of Sec. IV-A), providing the $N_1 \times N_2$ binary mask $K(x, y, t)$, which is the Activity Area for the first t frames. In

order to capture the temporal evolution of the activity using the kurtosis values, we estimate the Activity History Area (AHA) $K_H(x, y, t)$ as follows:

$$K_H(x,y,t) = \begin{cases} \tau, & if\ K(x,y,t) = 1; \\ max(0, K(x,y,t-1) - 1), & else. \end{cases} \quad (28)$$

If pixel (x, y) is active at time instant t (i.e. if $K(x, y, t) = 1$), its value in the AHA is that of the temporal window τ. If that pixel is inactive at time t, the AHA $K_H(x, y, t)$ obtains either the value zero, or its previous value $K_H(x, y, t-1)$ minus one, so it is either set to zero for a pixel that has not been active until time t, or its value is lower than that of currently active pixels (i.e. τ) if the pixel was active in the past. Thus, the frame pixels that were active in the past are weighted with lower values, while the most recently active pixels have higher AHA values. The result is, for both the MHI and AHA, a grayscale image with higher values in the regions that were active most recently, lower values in regions that were active in the past, and zero in regions that are constantly static. Since both the MHI and AHA are essentially time-weighted versions of the MEI and Activity Areas, they are expected to demonstrate similar degrees of robustness in the presence of additive noise. Indeed, in the experiments that follow, we show that the AHA remains more robust to noise than the MHI.

VI. Experiments

Experiments are conducted with a wide range of real sequences containing a variety of motions and themes, filmed both outdoors and indoors, with static and varying backgrounds. The use of MoG background modeling for the extraction of temporal templates is examined and shown to be less effective and practical than the MEI and Activity Area. Experiments with numerous videos demonstrate the increased robustness of the kurtosis-based approach to additive noise, compared to the difference-based method, both qualitatively and quantitatively. The effect of estimating temporal templates from fewer samples is examined by windowing the data with temporal masks of varying lengths, and comparing the results with those obtained from processing the entire video. Finally, challenging videos containing shadows or backgrounds with varying illumination caused by trees moving in the wind, rippling water, local occlusions, are also examined and it is shown that the proposed kurtosis-based approach provides better results in those cases as well.

A. MoG modeling for temporal templates

Modeling Accuracy: In this section we present some experimental results, to examine the usefulness of MoG modeling for the extraction of temporal activity templates. The background is modeled with a

MoG and active pixels are extracted by subtracting this background estimate from all frames, thresholding the result, and taking the union of the resulting foregrounds, as in Eq. (19), (20). Fig. 2 shows that this localizes the active pixels similarly to the MEIs and activity areas (Sec. VI-B, VI-C), but also contains many false alarms that originate from the rapid illumination changes in the video data, which cannot be followed by adaptive MoG models. This is a common drawback of adaptive MoG-based background modeling: it does not adapt rapidly enough to common illumination changes, because the accuracy of MoG modeling actually depends highly on tuning the model's parameters [39], [40], [11], which can be achieved only empirically, or via complex multivariate optimization. This limits its flexibility and general applicability, and renders it too computationally expensive and complex for practical applications.

Computation Time for MoG model based activity areas: In Table I we compare computation times for the MoG-based extraction of activity areas with that of kurtosis-based activity areas and MEIs (Sec. IV, II-C). We use Matlab 7.0 on a DCPU Pentium IV PC at 3.40GHz. The computation time for the Activity Areas and the MEI is essentially the same, as it differs by fractions of a second. Table I shows that, for all the videos, the MoG-based method has significantly higher computational times. It should be noted that the different computation times between videos (and not methods) are due to the difference in the dimensions of the frames being processed and the length of the sequences.

From the analysis of Sec. VI-A, we conclude that the MoG does not always lead to activity areas as accurate as those extracted from the simpler methods, since the estimated distribution model does not always adapt to scene changes quickly enough. It also requires significantly more computational time than the other approaches, because of the background modeling stage. Consequently, in our problem, namely that of isolating active pixels, we do not use the MoG for the extraction of active pixel areas. We focus on the proposed kurtosis-based approach and the difference-based temporal templates (see Sec. IV, VI-C), which achieve more consistent and accurate results at a lower computational cost.

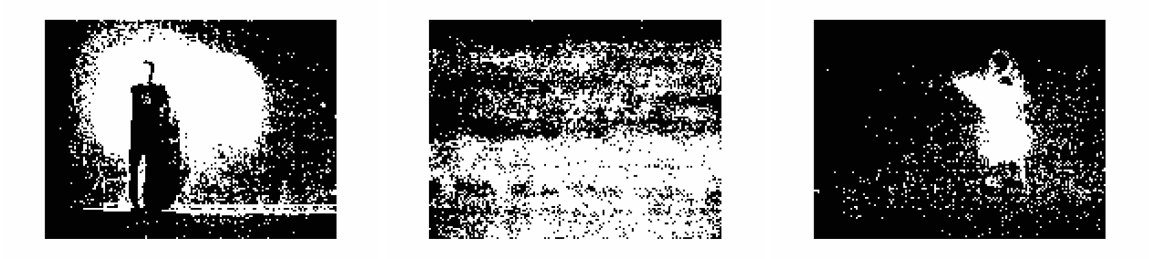


Fig. 2. MoG-based temporal templates for Wave person 23, Jog person 11-d1, Box person 13-d3.

TABLE I

COMPUTATION TIMES (SEC) FOR MOG, MEI, KURTOSIS-BASED ACTIVITY AREAS.

Template/Video	Hoop	Kid Plane 1	Kid Plane 2	Jog 11-d1	Jog 11-d2	Run 11-d1	Run 11-d2	Run 23
MoG	111	450	321	26	28	32	30	35
MEI/Act. Area	32	115	89	6	7	8	7.8	8.5
Template/Video	Box 4	Box 13-d1	Box 13-d3	Clap 4	Clap 13-d1	Clap 13-d4	Walk	Wave
MoG	23	31	28	29	26	28	38	32
MEI/Act. Area	2	2.3	1.8	2	1.6	1.7	3	2.5

B. Kids videos

In this experiment, the performance of the MEI against the Activity areas in the presence of noise is examined for videos filmed outdoors, showing kids shooting hoops and throwing toy airplanes. Representative frames of the videos masked by the Activity Areas are shown in Fig. 3, and the entire videos, masked by both the MEI and Activity Area can be found at http://mklab.iti.gr/robust-temporal-templates. In the presence of additive noise, the MEIs degrade significantly, making it difficult to separate the truly active pixels from the false alarms, whereas the activity areas are nearly unaffected, as seen in Fig. 4. Fig. 5 shows how the mean absolute error increases for all temporal templates under increasing noise. This error is the mean of the absolute difference between the AA-MEI or AHA-MHI extracted in the absence of noise, and in the presence of increasing noise: it is clear that the MEIs and MHIs are much more sensitive to additive noise than the AAs and AHAs, as expected from the qualitative results.

C. Human Motions

We conduct numerous experiments with videos of people performing various activities, such as running, jogging, clapping, which can be viewed at http://mklab.iti.gr/robust-temporal-templates. For activities that appear in more than one videos, we present qualitative results for one of them for reasons of space, but quantitative results are shown for all cases in Sec. VI-D. Some examples of Activity Areas superposed on video frames are shown in Fig. 6, and the entire videos, masked by the MEI and the Activity Areas, can be found at http://mklab.iti.gr/robust-temporal-templates. As in Sec. VI-B, the Activity Areas prove to be more robust to additive noise than the MEIs. This can be seen in Fig. 7, where the noise introduces many false alarms in the MEIs but almost none in the Activity Areas. The noise in the MEIs appears both as small white dots in the static pixels, but also degrades the overall shape of the MEIs. Thus, morphological

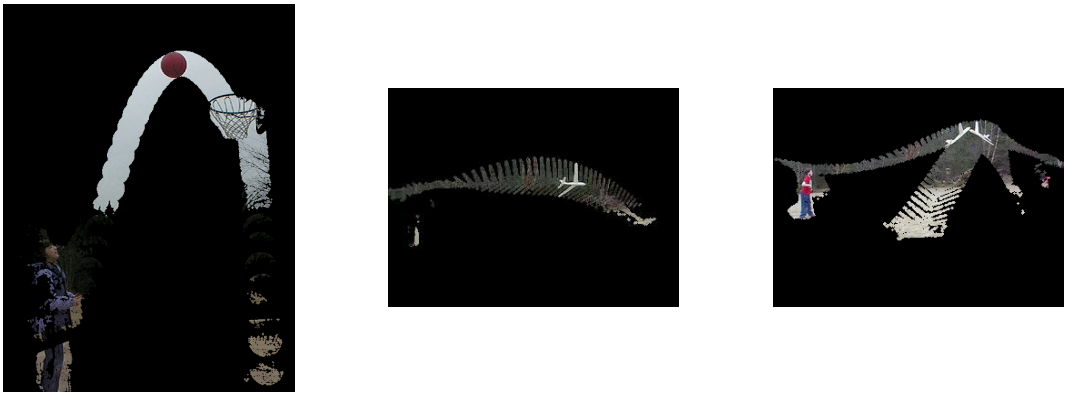


Fig. 3. Video frames masked by Activity Areas for Hoop, Kid Plane 1, Kid Plane 2.

processing of the noisy MEIs would eliminate the false alarms, but also “erase” significant parts of the contour, which are characteristic of the activity taking place. The degradation introduced by the additive noise is particularly evident in the plots of the mean absolute error, in Fig. 8. Here, we see that for a wide range of motions, the error in the MEIs and MHIs increases significantly, whereas the Activity Areas and Activity History Areas are not affected much by the noise.

D. Quantitative Comparison

In this section, we provide quantitative results comparing the proposed kurtosis-based approach for the estimation of the Activity Area and Activity History Area with the difference-based method of finding the MEI and MHI in the presence of noise. The mean difference between all noisy templates and the original one, estimated under no noise, is the mean absolute error plotted in Figs 5, 8. Tables II and III show the mean of this error for each video: it is essentially the mean of the absolute errors plotted in Fig. 5, 8. The Activity Area and Activity History Area, remain less affected by the additive noise than the MEI and MHI for the same data sets, so they are considered a more reliable measure of pixel activity, which can be obtained for a very similar computational cost.

E. Effect of Varying Temporal Window

In this work we have made the assumption that the image statistics do not change from frame to frame. In practice however, the kurtosis is estimated using arithmetic means over time (video frames), so using a smaller amount of frames translates into fewer samples, which may lead to small deviations from the

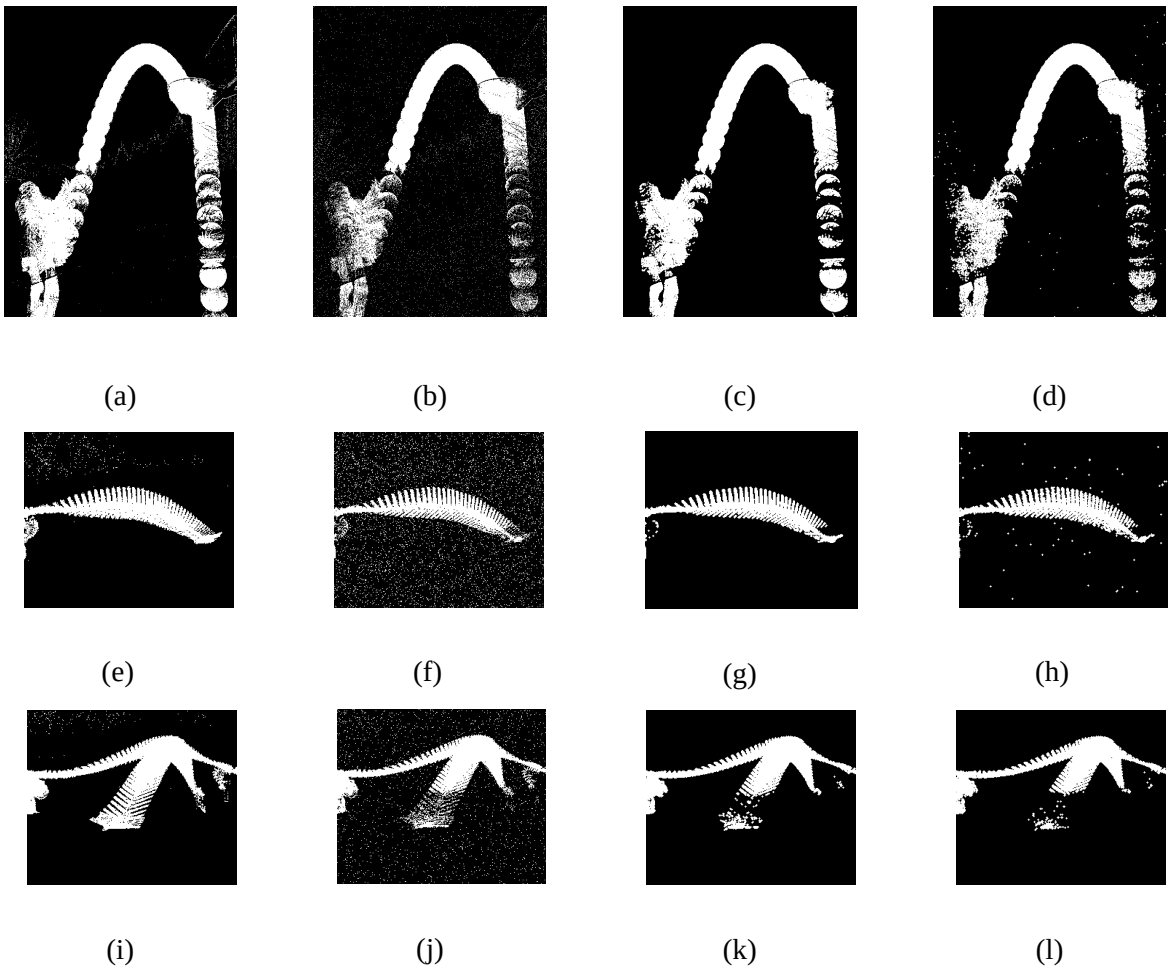


Fig. 4. Effect of noise on temporal templates. In each pair of figures figures, the first is for noiseless, the second for noisy data. Hoop: (a),(b) MEI (c),(d) AA. Kid Plane 1: (e),(f) MEI, (g),(h) AA. Kid Plane 2: (i),(j) MEI (k),(l) AA.

true expected values. The difference-based method does not compute statistics, as it simply estimates the union of binary masks extracted from pairs of frames, so it is not affected by the number of frames used. In order to test the proposed method's robustness to using fewer frames, experiments are performed by estimating the MEI and Activity Areas with temporal windows of varying length, ranging from as few as 2 frames to the entire sequence length. The results for each temporal template and for each window size are compared with the results obtained by using the entire video sequence, and the resulting average absolute differences are plotted (on the same scale) in Fig. 9 for representative videos. The Activity Area estimated from fewer frames deviates from its estimate using the entire video, however this deviation decreases significantly as more frames are used. As can be seen from Fig. 9, the deviation of the Activity

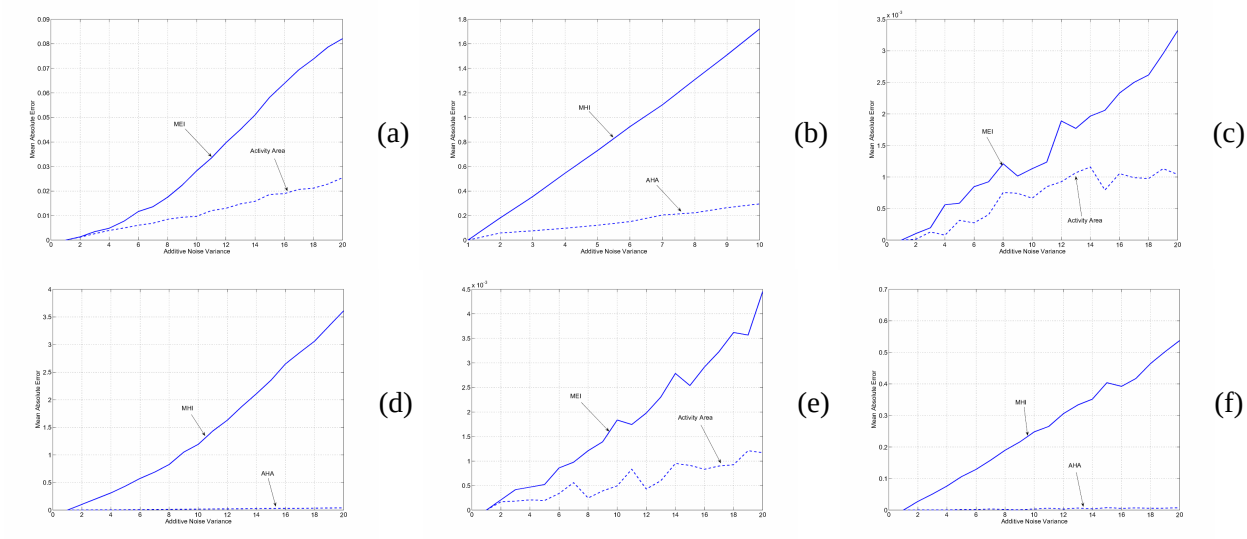


Fig. 5. Mean absolute error under increasing noise. Top row, MEI and Activity Area. Bottom row, MHI and Activity History Area. (a), (d) Hoop, (b), (e) Kid plane 1, (c), (f) Kid plane 2.

Fig. 6. Video frames masked by Activity Areas for two boxing, one clapping, one handwaving videos.

Area computed using fewer frames from that computed by using all frames is generally not very high, even for very short temporal windows: temporal windows of 2 or 3 frames are very small, and yet still provide a reliable template of the Activity Area in some cases, e.g. for the Basketball Hoop, Box or Pool videos, in Fig. 10. In other videos, small errors appear in the background when much fewer frames than the entire sequence are used, such as the Trees and Kid Flying Plane sequences, which contain over 100 frames. This is not unexpected, since the kurtosis is estimated from the arithmetic average over the video frames, which becomes more reliable as the number of frames used increases. Nevertheless, the active pixels are still correctly extracted, and Activity Areas with almost no errors can still be obtained by using longer subsequences, but not the entire video.

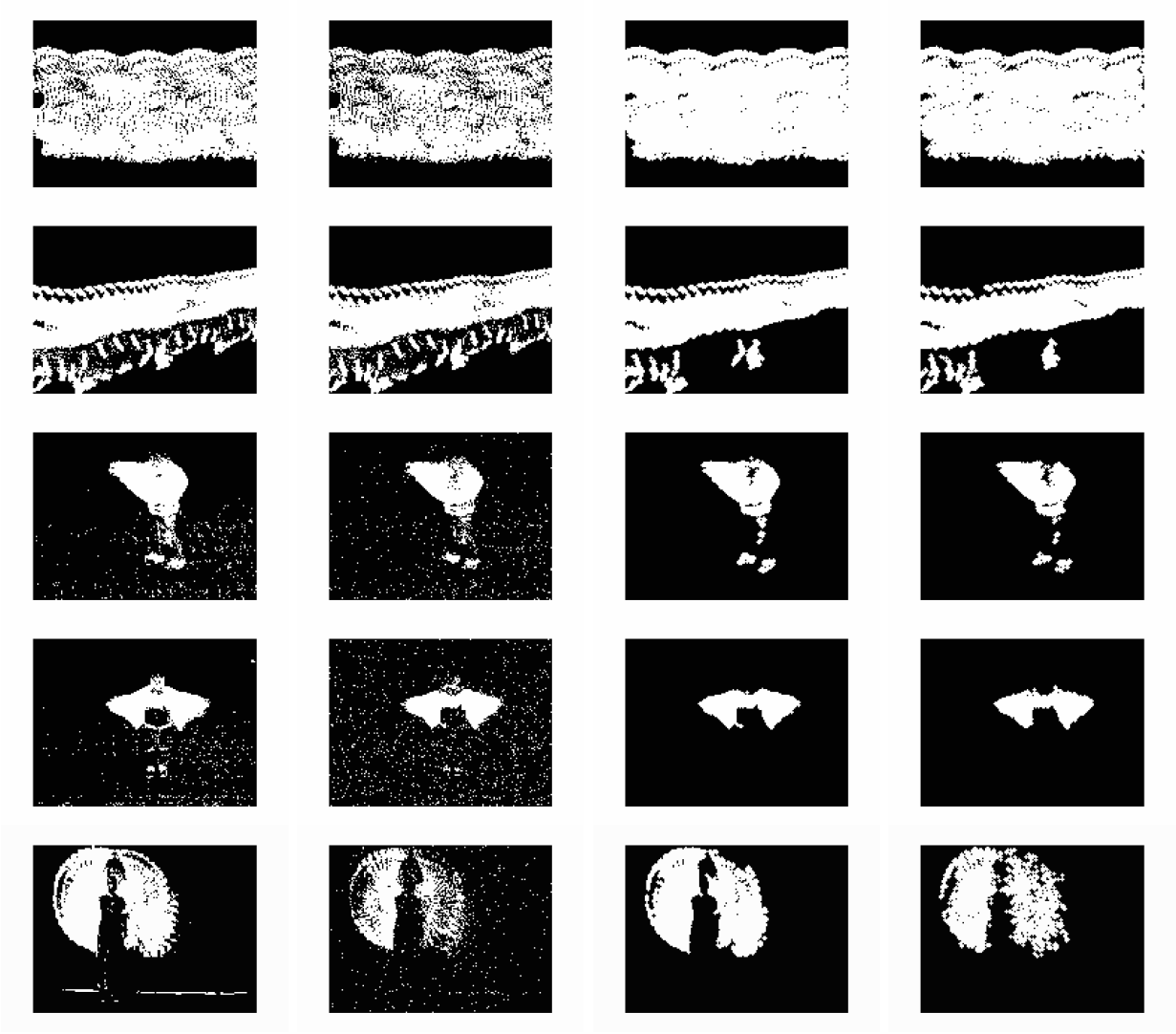


Fig. 7. Effect of noise on temporal templates for various human activities. Each line has a MEI for noiseless data, a MEI for noisy data, an AA for noiseless data, an AA for noisy data.

F. Videos with Shadows

In this section a set of experiments with videos containing shadows takes place. The temporal templates are extracted as before, and postprocessing takes place in order to remove the pixels corresponding to shadows based on color information [41]. The HSV color space is used as it matches human perception of color [42], and has shown to be a shadow invariant space [43], although other color spaces can also be used [44]. The videos examined are from games of pool, where the pool ball, the stick and the hand of the player create a shadow on the table. Fig. 11 shows that the MEI and Activity Area results

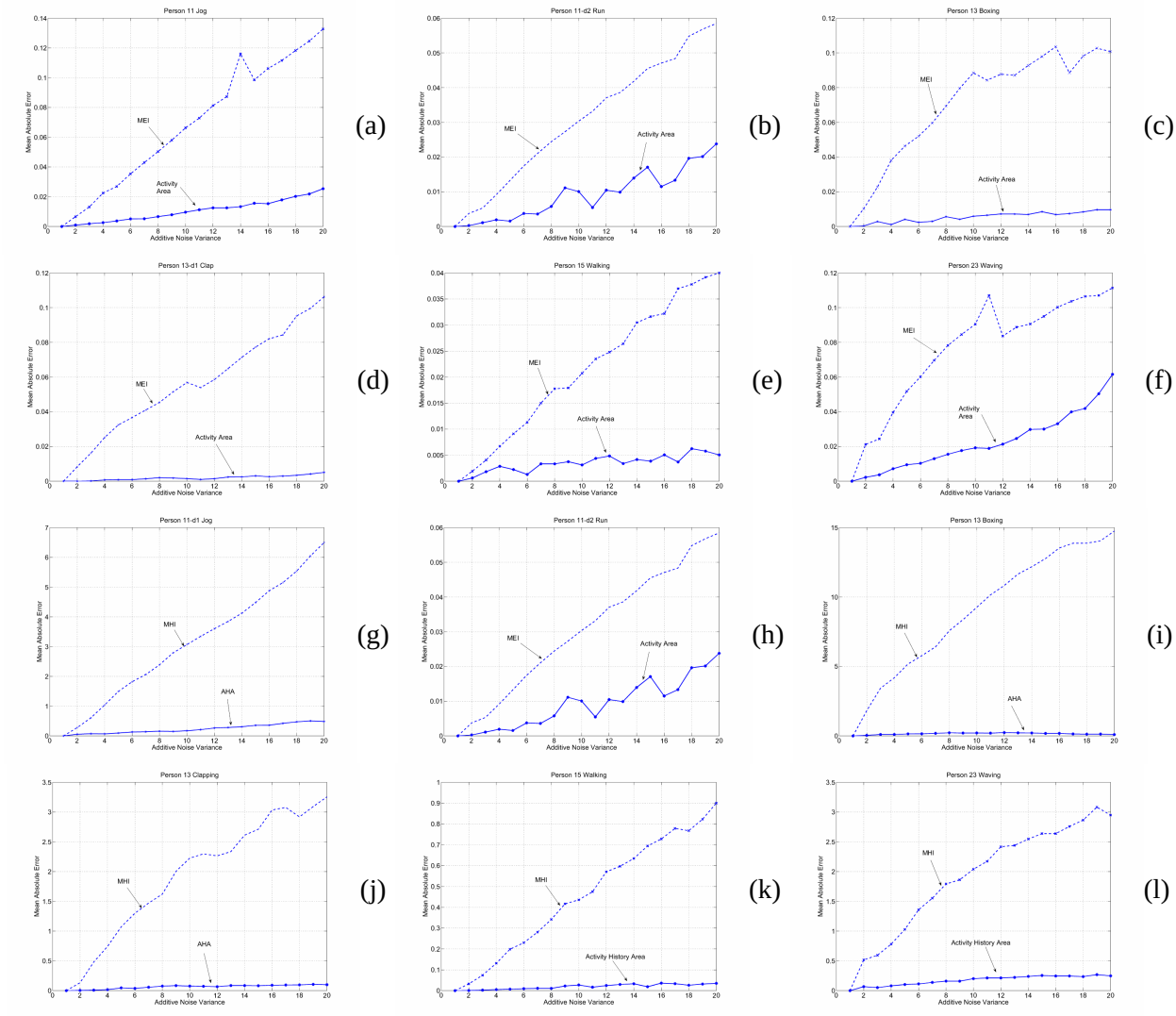


Fig. 8. Mean absolute error under increasing noise. MEI and AA: (a) Jog, (b) Run, (c) Box, (d) Clap (e) Walk (f) Wave. MHI and AHA: (g) Jog, (h) Run, (i) Box, (j) Clap (k) Walk (l) Wave

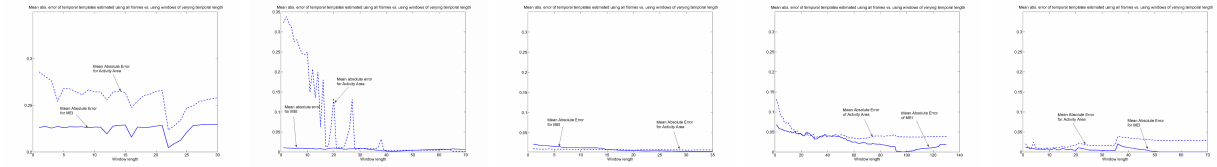


Fig. 9. Mean absolute error from varying window length. Basketball hoop, Kid flying plane, Trees, Pool topspin.

are comparable for these videos, although the MEI is slightly noisier, as can also be seen from the videos on http://mklab.iti.gr/robust-temporal-templates. Color based removal of the shadows combined

TABLE II

MEAN ABSOLUTE ERROR OF MEI AND ACTIVITY AREA.

Template/Video	Hoop	Kid Plane 1	Kid Plane 2	Jog 11-d1	Jog 11-d2	Run 11-d1	Run 11-d2	Run 23
MEI	0.1256	0.0325	0.0938	0.0686	0.0403	0.0677	0.0307	0.0225
Activity Area	0.0279	0.007	0.0164	0.0105	0.0092	0.0134	0.0091	0.0069
Template/Video	Box 4	Box 13-d1	Box 13-d3	Clap 4	Clap 13-d1	Clap 13-d4	Walk	Wave
MEI	0.0617	0.0705	0.0954	0.0694	0.0554	0.0381	0.0063	0.0685
Activity Area	0.0042	0.0054	0.0084	0.0013	0.0019	0.0020	0.0046	0.0170

TABLE III

MEAN ABSOLUTE ERROR OF MHI AND AHA.

Template/Video	Hoop	Kid Plane 1	Kid Plane 2	Jog 11-d1	Jog 11-d2	Run 11-d1	Run 11-d2	Run 23
MHI	3.0458	2.5040	4.3244	3.1539	1.9253	0.4931	0.8216	2.8312
AHA	0.0926	0.0655	0.0700	0.2365	0.1225	0.0550	0.0775	0.1778
Template/Video	Box 4	Box 13-d1	Box 13-d3	Clap 4	Clap 13-d1	Clap 13-d4	Walk	Wave
MHI	1.7635	8.9850	4.4636	2.3073	1.9328	1.9953	1.3789	5.1270
AHA	0.0353	0.1561	0.2691	0.0575	0.0643	0.1201	0.0728	0.2575

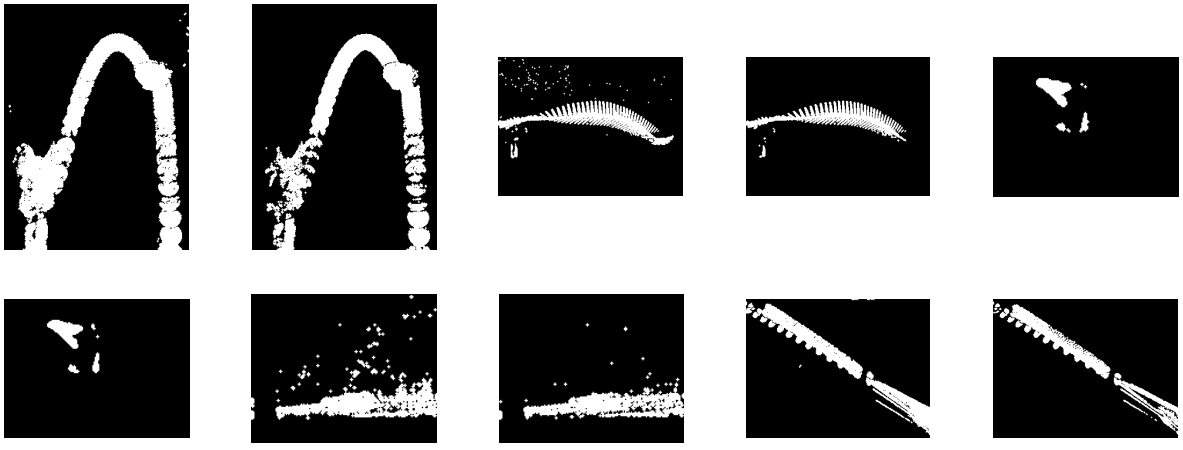


Fig. 10. AA for: Basketball hoop, windows of 2, 10 frames. Kid flying plane, windows of 24, 70 frames. Boxing, windows of 5, 100 frames. Trees, windows of 15, 50 frames. Pool, windows of 3, 30 frames.

with the Activity Area leads to the results of Fig. 11(d), (h), where it can be seen that the shadow pixels have been effectively removed, without affecting the truly active pixels. Videos masked by the Activity Area and MEI after shadow removal can be seen on http://mklab.iti.gr/robust-temporal-templates. Many different kinds of post-processing can be applied for shadow removal, but detailed investigation into them is beyond the scope of this paper, as they depend on the application and information available, requiring e.g. information about the direction of the light source, 3D geometry of the background [45], or assumptions about the direction of the shadow [46].

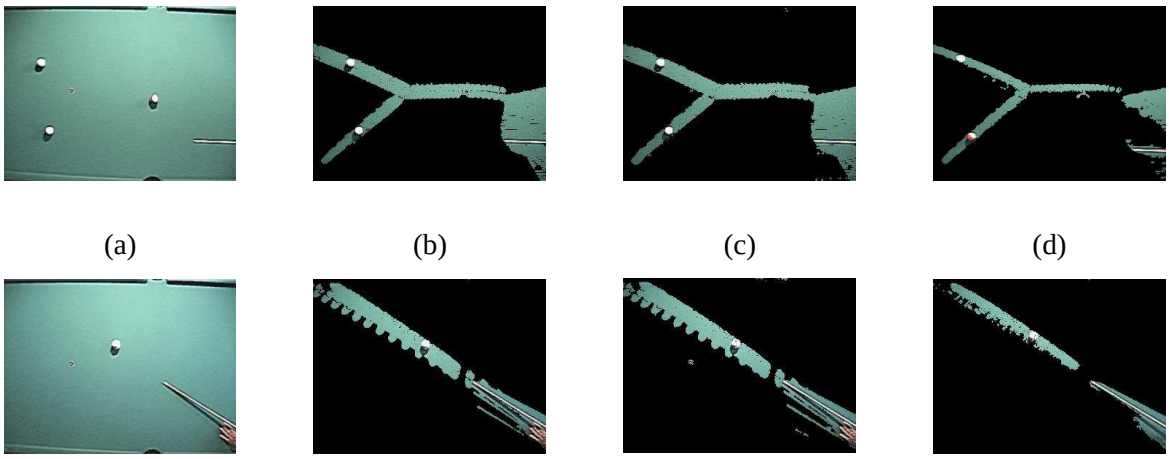


Fig. 11. Curve ball. (a) frame 20 (b) MEI (c) Activity Area (d) Activity Area without shadow. Topspin: (e) frame 50 (f) MEI (g) Activity Area (h) Activity Area without shadow.

G. Videos with Varying Backgrounds and Occlusions

The proposed kurtosis-based method is expected to remain robust to deviations of the background illumination changes from the single Gaussian model, due to the robustness of the kurtosis metric. Its global nature is also expected to make it robust to local occlusions and illumination changes. Experiments with challenging sequences, containing backgrounds with tree leaves moving in the wind, local occlusions, crowds in stadiums, rippling water, take place to show that, indeed, in those cases the Activity Areas remain more accurate than the MEIs (see http://mklab.iti.gr/robust-temporal-templates). The videos containing trees moving in the wind contain significant background motion, which is not mistaken for the actual foreground motion by the kurtosis-based approach, although it introduces several artifacts in the MEIs. These videos also contain local occlusions, however both methods remain unaffected by them,

due to their global nature. In the video of the polevault jumper, there are small motions in the crowd. The MEI for this case is significantly degraded, whereas the Activity Area correctly captures the active pixels. Finally, the video with the polar bear contains motion of the water, which is higher in the front of the bear's island. As before, the MEI erroneously incorporates many of the moving background pixels, and the Activity Area proves much more robust, isolating only the truly active pixels, but also a region of the water with higher activity. Videos containing the original videos, the videos masked by the MEI and the Activity Area, can be seen in http://mklab.iti.gr/robust-temporal-templates, where it can be seen that the Activity Areas indeed significantly outperform the MEIs in more challenging cases.

VII. Conclusions, Future Work

We have presented a novel, robust, computationally efficient approach for the characterization of pixels in a video sequence as active. The proposed method is based on the processing of higher order statistics of the pixels' illumination changes over time, and is compared with often-used difference-based solutions. Unlike existing background removal techniques, it does not require prior knowledge, a training stage, and has a low computational cost. The kurtosis is used as an indicator of pixel activity, as its value is theoretically zero when a pixel is static and the measurement noise follows a Gaussian distribution over time. Theoretical analysis demonstrates the robustness of the proposed approach to additive noise, and the sensitivity of difference-based methods to it. The deviation of additive noise from a Gaussian distribution when the background is modeled by MoGs is shown to affect the kurtosis very little. MoGs are shown to be computationally more costly and sensitive to correct initialization of model parameters, so they are not included in the comparison of methods for temporal templates. Extensive experiments demonstrate the effect of additive noise on the Activity Areas generated by the proposed, kurtosis-based method, and on the MEIs, as well as its effect on the Activity History Areas and MHIs. Both qualitative and quantitative results show that higher-order statistics provide a more reliable and robust measure of pixel activity. By applying temporal windows of varying size to obtain the kurtosis-based Activity Areas, it is shown that the method remains robust to using smaller amounts of data. The issue of the presence of shadows is also addressed via post-processing, which successfully removes the shadow pixels, while retaining the truly active pixels in the temporal templates. Finally, challenging videos containing moving backgrounds and occlusions are examined, and it is shown that the kurtosis-based approach can provide more accurate results than the difference-based one, even under realistic and difficult conditions.

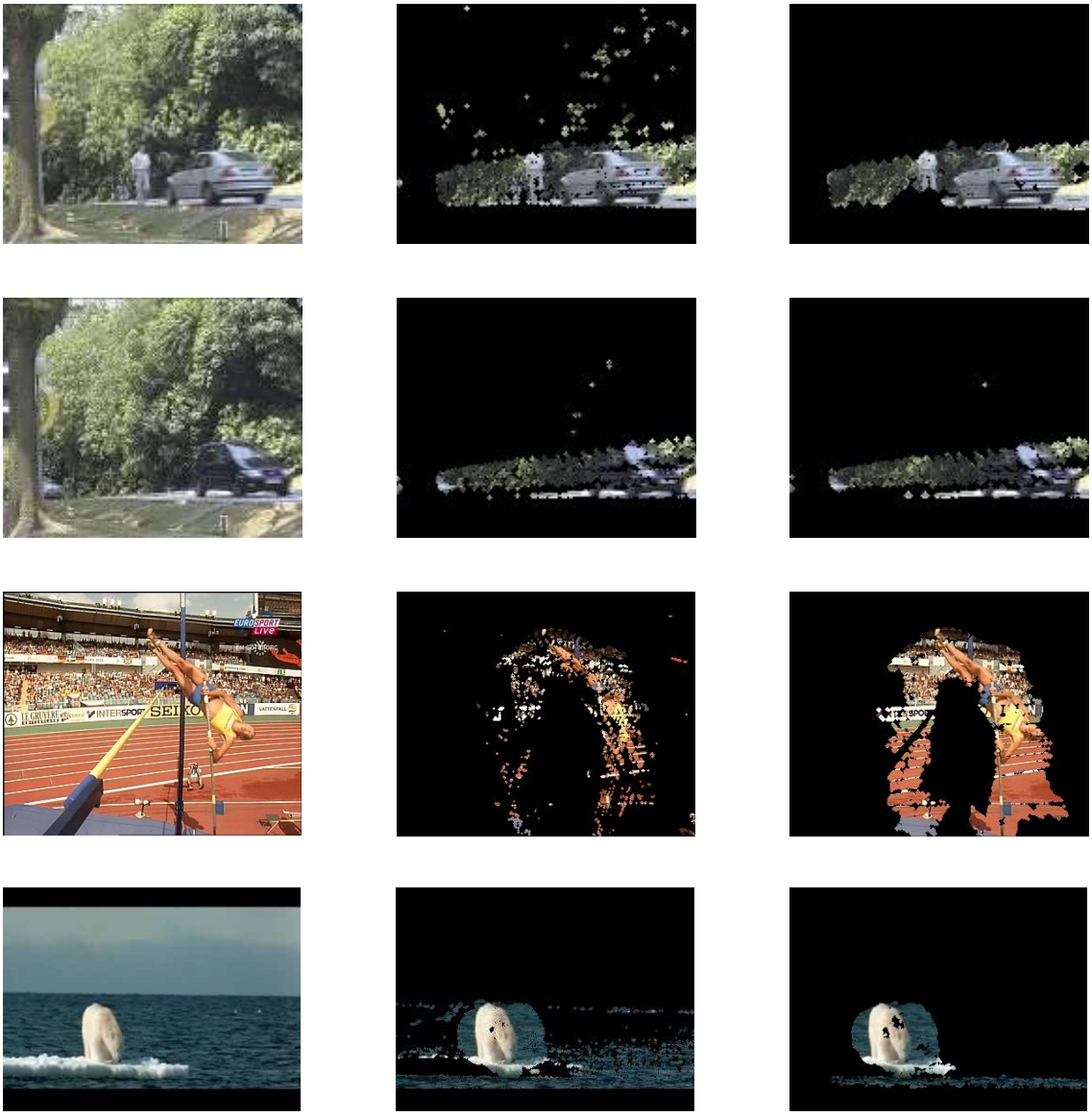


Fig. 12. Videos with moving tree leaves in the background, a polevault jumper and a bear, with non-static backgrounds. Each row has a ideo frame, video frame masked by MEI, video frame masked by AA. In all cases the AA mask contains fewer artifacts.

APPENDIX A
DERIVATION OF K_{dI}

The kurtosis of illumination changes given by Eq (10) is derived below.

$$K_{dI} = E[(I(x,y,t) - I(x,y,t-1))^4] - 3(E[(I(x,y,t) - I(x,y,t-1))^2])^2, \quad (29)$$

where

$$E[(I(x,y,t)-I(x,y,t-1))^4] = E[I^4(x,y,t)+I^4(x,y,t-1)+2I^2(x,y,t)I^2(x,y,t-1)$$
$$+4I^2(x,y,t)I^2(x,y,t-1)-4I(x,y,t)I(x,y,t-1)(I^2(x,y,t)+I^2(x,y,t-1)]$$
$$= E[I^4(x,y,t)]+E[I^4(x,y,t-1)]+2E[I^2(x,y,t)I^2(x,y,t-1)]$$
$$+4E[I^2(x,y,t)I^2(x,y,t-1)]-4E[I^3(x,y,t)I(x,y,t-1)]-4E[I(x,y,t)I^3(x,y,t-1)]$$
$$= 2m_4+2E[I^2(x,y,t)I^2(x,y,t-1)]+4E[I^2(x,y,t)]E[I^2(x,y,t-1)]$$
$$-4E[I^3(x,y,t)I(x,y,t-1)]-4E[(I(x,y,t)I^3(x,y,t-1)], \quad (30)$$

where we have taken into account the assumptions of Sec. IV-A, namely that the k_{th} order moments of each frame's value remain the same over successive frames. The notation of Eq. (3) is also used here. Also:

$$(E[(I(x,y,t)-I(x,y,t-1))^2])^2 = (E[I^2(x,y,t)])^2+(E[I^2(x,y,t-1)])^2$$
$$+2E[I^2(x,y,t)]E[I^2(x,y,t-1)]+4E^2[I(x,y,t)I(x,y,t-1)]$$
$$-4E[I(x,y,t)I(x,y,t-1)]E[I^2(x,y,t)]-4E[I(x,y,t)I(x,y,t-1)]E[I^2(x,y,t-1)] \quad (31)$$

By plugging Eqs. (30), (31) into Eq. (29), we get:

$$K_{dI} = 2K_I+6(E[I^2(x,y,t)I^2(x,y,t-1)]-E[I^2(x,y,t)]E[I^2(x,y,t-1)])-4E[I^3(x,y,t-1)I(x,y,t)]$$
$$-4E[I^3(x,y,t)I(x,y,t-1)]-12(E[I(x,y,t)I(x,y,t-1)])^2+12E[I(x,y,t)I(x,y,t-1)]E[I^2(x,y,t)]$$
$$+12E[I(x,y,t)I(x,y,t-1)]E[I^2(x,y,t-1)] = 2K_I+K_c, \quad (32)$$

where K_I is the kurtosis of each frame's luminance values $I(x,y,t)$, considered to be the same for all frames and by K_c we denote the cross-correlation of successive frames' illumination values (and their higher powers). If we consider that illumination values from different frames and their powers are independent from each other, as well as that their higher order moments are the same, Eq. (31) becomes $K_{dI} = 2K_I - 8m_1m_2 - 12m_1^4 + 24m1^2m_2$.

APPENDIX B
DERIVATION OF KURTOSIS FOR MOG

The derivation of the kurtosis for the general case of a MoG is provided below. The pdf for a MoG is $f_X(x) = \sum_{i=1}^{N} w_i f_i(x)$, where f_i are Gaussian distributions with mean μ_i and variance σ_i^2. The higher

order moments of X from Eq. (16) are $E[X^k] = \sum_{i=1}^{N} w_i E_{fi}[X^k]$, with $E_{fi}[X^k]$ the k^{th} order moments of Gaussian f_i. The kurtosis of X is $K_X = E[X^4] - 3E^2[X^2]$, where $E[X^4] = \sum_{i=1}^{N} w_i E_{fi}[X^4]$ and

$$E^2[X^2] = (\sum_{i=1}^{N} w_i E_{fi}[X^2])^2 = \sum_{i=1}^{N} w_i^2 E_{fi}^2[X^2] + \sum_{i \neq j} w_i w_j E_{fi}[X^2] E_{fj}[X^2]. \quad (33)$$

The kurtosis of X is then given by:

$$\begin{aligned} K_X &= \sum_{i=1}^{N} w_i E_{fi}[X^4] - 3\sum_{i=1}^{N} w_i^2 E_{fi}^2[X^2] - 3\sum_{i \neq j} w_i w_j E_{fi}[X^2] E_{fj}[X^2] \\ &= \sum_{i=1}^{N} w_i (E_{fi}[X^4] - 3E_{fi}^2[X^2] - 3(w_i - 1)E_{fi}^2[X^2]) - 3\sum_{i \neq j} w_i w_j E_{fi}[X^2] E_{fj}[X^2] \\ &= \sum_{i=1}^{N} w_i K_{Xi} - 3w_i(w_i - 1)E_{fi}^2[X^2] - 3\sum_{i \neq j} w_i w_j E_{fi}[X^2] E_{fj}[X^2] \\ &= 3\Big(\sum_{i=1}^{N} w_i(1 - w_i)E_{fi}^2[X^2]) - \sum_{i \neq j} w_i w_j E_{fi}[X^2] E_{fj}[X^2]\Big), \end{aligned} \quad (34)$$

where $K_{Xi} = E_{fi}[X^4] - 3E_{fi}^2[X^2] = 0$, since the individual mixtures f_i are Gaussian and have zero kurtosis. This quantity, representing the kurtosis of data that follows a MoG distribution, obtains small values, as the weights are less than one, and become very small when multiplied with each other.

APPENDIX C
ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Communitys Seventh Framework Programme FP7/2007-2013 under grant agreement FP7-214306 - JUMAS, from FP6 under contract number 045547- VidiVideo.

REFERENCES

[1] A. Bobick and J. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257–267, 2002.

[2] G. Bradski and J. Davis.

[3] T. Ogata, J. Tan, and S. Ishikawa, "High-speed human motion recognition based on a motion history image and an eigenspace," *IEICE Transactions on Information and Systems*, no. 1, pp. 281–289, 2006.

[4] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *The Tenth IEEE International Conference on Computer Vision (ICCV'05)*, 2005, pp. 1395–1402.

[5] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2247–2253, December 2007.

[6] G. Xu, Y. Ma, H. Zhang, and S. Yang, "Motion based event recognition using HMM," in *Proc. 16th International Conference on Pattern Recognition (ICPR'02)*, vol. 2, 2002.

[7] R. Babu, B. Anantharaman, K. Ramakrishnan, and S. Srinivasan, "Compressed domain action classification using hmm," *Pattern Recognition Letters*, vol. 23, no. 10, pp. 1203–1213, 2002.

[8] R. Fablet and P. Bouthemy, "Motion recognition using nonparametric image motion models estimated from temporal and multiscale co-occurrence statistics," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 12, pp. 1619 – 1624, 2003.

[9] W. Hu, X. Xiao, Z. Fu, D. Xie, T. Tan, and S. Maybank, "A system for learning statistical motion patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 9, pp. 1450–1464, 2006.

[10] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," in *Proc. IEEE Conf. Computer Vision Patt. Recog.*, June 1999, pp. 246–252.

[11] Z. Zivkovic and F. van der Heijden, "Efficient adaptive density estimation per image pixel for the task of background subtraction," *Pattern Recogn. Lett.*, vol. 27, no. 7, pp. 773–780, 2006.

[12] T. Aach, A. Kaup, and R. Mester, "Statistical model-based change detection in moving video," *Signal Processing*, vol. 31, no. 2, pp. 165–180, 1993.

[13] J. Pilet, C. Strecha, and P. Fua, "Making background subtraction robust to sudden illumination changes," in *European Conference on Computer Vision, ECCV 2008*, 2008, pp. 567–580.

[14] A. Mittal and N. Paragios, "Motion-based background subtraction using adaptive kernel density estimation," in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. IEEE Computer Society Conference on*, vol. 2, June 2004, pp. 302–309.

[15] D. H. A. Elgammal and L. Davis, "Nonparametric model for background subtraction," in *Proc. European Conf. Computer Vision*, June 2000, pp. 751–767.

[16] M. Irani, B. Rousso, and S. Peleg, "Computing occluding and transparent motions," *International Journal of Computer Vision*, vol. 12, pp. 5–16, 1994.

[17] P. Bouthemy, C. Hardouin, G. Piriou, and J. Yao, "Mixed-state auto-models and motion texture modeling," *J. Math. Imaging Vis.*, vol. 25, no. 3, pp. 387–402, 2006.

[18] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, "Pfinder: real-time tracking of the human body," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 780–785, 1997.

[19] E. Sayrol, A. Gasull, and J. R. Fonollosa, "Motion estimation using higher order statistics," *IEEE Transactions on Image Processing*, vol. 5, no. 6, pp. 1077–1084, June 1996.

[20] K.-P. Karmann, A. von Brandt, and R. Gerl, *Moving Object Segmentation Based on Adaptive Reference Images*. Elsevier Science, 1990.

[21] M. Hassouni, H. Cherifi, and D. Aboutajdine, "Hos-based image sequence noise removal," *IEEE Transactions on Image Processing*, vol. 15, no. 3, pp. 572–581, 2006.

[22] J. Anderson and G. Giannakis, "Image motion estimation algorithms using cumulants," *IEEE Transactions on Image Processing*, vol. 4, no. 3, pp. 346 – 357, 1995.

[23] J. Tugnait, "Time delay estimation with unknown spatially correlated gaussian noise," *IEEE Transactions on Signal Processing*, vol. 41, no. 2, pp. 549 – 558, 1993.

[24] J. Duncan and T. Chou, "On the detection of motion and the computation of optical flow," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 3, pp. 345–352, March 1992.

[25] M. Proesmans, L. J. V. Gool, E. J. Pauwels, and A. Oosterlinck, "Determination of optical flow and its discontinuities

using non-linear diffusion," in *ECCV '94: Proceedings of the Third European Conference-Volume II on Computer Vision*. London, UK: Springer-Verlag, 1994, pp. 295–304.

[26] 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*, June 1992, pp. 236–242.

[27] M. Welling, "Robust higher order statistics," in *Tenth International Workshop on Artificial Intelligence and Statistics*, Barbados, Jan. 2005, pp. 405–412.

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

[29] Casella and Berger, *Statistical Inference*. Duxbury, 2002.

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

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

[32] 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.

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

[34] P. A. Delaney, "Signal detection using third-order moments," *Circuits Systems Signal Process*, vol. 13, no. 4, pp. 481–496, 1994.

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

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

[37] J. R. Blum and J. Rosenblatt, "On random sampling from a stochastic process," *The Annals of Mathematical Statistics*, vol. 35, no. 4, pp. 1713–1717, Dec. 1964.

[38] J. K. Tugnalt, "On time delay estimation with unknown spatially correlated gaussian noise using the 4th-order cumulants and cross cumulants," *IEEE Transactions on Signal Processing*, vol. SP-39, no. 6, pp. 1258–1267, 1991.

[39] A. Penalver, J. Saez, and F. Escolano, "An entropy maximization approach to optimal model selection in gaussian mixtures," in *Progress in Pattern Recognition, Speech and Image Analysis*, vol. 2905, 2003, pp. 432–439.

[40] F. C. X. Gao, T. Boult and V. Ramesh, "Error analysis of background adaption," in *Proceedings of IEEE Computer Vision and Pattern Recognition, CVPR*, 2000, pp. 503–510.

[41] J. Geusebroek, R. van den Boomgaard, A. Smeulders, and H. Geerts, "Computing occluding and transparent motions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 12, pp. 1338–1350, 2001.

[42] N. Herodotou, K. Plataniotis, and A. Venetsanopoulos, "A color segmentation scheme for object-based video coding," in *Advances in Digital Filtering and Signal Processing, 1998 IEEE Symposium on*, 1998, pp. 25 – 29.

[43] R. Cucchiara, C. Grana, M. Piccardi, A. Prati, and S. Sirotti, "Improving shadow suppression in moving object detection with HSV color information," in *Intelligent Transportation Systems, 2001. Proceedings. 2001 IEEE*, 2001, pp. 334–339.

[44] E. Salvador, P. Green, and T. Ebrahimi, "Shadow identification and classification using invariant color models," in *Proceedings of ICASSP 01*, vol. 3, 2001, pp. 1545–1548.

[45] D. Koller, K. Danilidis, and H.-H. Nagel, "Model-based object tracking in monocular image sequences of road traffic scenes," *International Journal of Computer Vision*, vol. 10, no. 3, p. 257281, 1993.

[46] C. Chang, W. Hu, J. Hsieh, and Y. Chen, "Shadow elimination for effective moving object detection with gaussian models," in *Pattern Recognition, International Conference on*, vol. 23, 2002.