Detection of ruptures in spatial relationships in video sequences

Abdalbassir Abou-Elailah[1], Valerie Gouet-Brunet[1] and Isabelle Bloch[2]
[1] *Université Paris-Est, IGN, SRIG, MATIS, 73 avenue de Paris, 94160 Saint Mandé, France*
[2] *Institut Mines-Telecom, Telecom ParisTech, CNRS LTCI, France*
{*Abd-Al-Bassir.Abou-El-Ailah, valerie.gouet*}*@ign.fr, isabelle.bloch@telecom-paristech.fr*

Keywords:
Spatial relationships, Angle histogram, Distances, Fuzzy object representation, Detection of ruptures.

Abstract:
The purpose of this work is to detect strong changes in spatial relationships between objects in video sequences, with a limited knowledge on the objects. First, a fuzzy representation of the objects is proposed based on low-level generic primitives. Furthermore, angle and distance histograms are used as examples to model the spatial relationships between two objects. Then, we estimate the distances between different angle or distance histograms during time. By analyzing the evolution of the spatial relationships during time, ruptures are detected in this evolution. Experimental results show that the proposed method can efficiently detect the ruptures in the spatial relationships, exploiting only low-level primitives. This constitutes a promising step towards event detection in videos, with few a priori models on the objects.

1 INTRODUCTION

In the literature, there are many intelligent video surveillance systems, and each system is dedicated to a specific application, such as sport match analysis, people counting, analysis of personal movements in public shops, behavior recognition in urban environments, drowning detection in swimming pools, etc.[1] The VSAM project (Visam, 1997) was probably one of the first projects dedicated to surveillance from video sequences. The goal of ICONS project (Icons, 2000) was to recognize the incidents in video surveillance sequences. The goal of the three projects ADVISOR (Advisor, 2000), ETISEO (Etiseo, 2004) and CareTracker (Caretaker, 2006) was to analyze record streaming video, in order to recognize events in urban areas and to evaluate scene understanding. The AVITRACK project (Avitrackr, 2004) was applied to the monitoring of airport runways, while the BEWARE project (Beware, 2007) aimed to use dense camera networks for monitoring transport areas (railway stations, metro).

In this context, an increasing attention is paid to "event" detection. In (Piciarelli et al., 2008), an approach is proposed to detect anomalous events based on learning 2-D trajectories. In (Saleemi et al., 2009), a probabilistic model of scene dynamics is proposed for applications such as anomaly detection and improvement of foreground detection. For crowded scenes, tracking moving objects becomes very difficult due to the large number of persons and background clutter. There are many approaches proposed in the literature for abnormal event detection, based on spatio-temporal features. In (Jiang et al., 2009), an unsupervised approach is proposed based on motion contextual anomaly of crowd scenes. In (Mehran et al., 2009), a social force model is used for abnormal crowd behavior detection. In (Cong et al., 2013b), an abnormal event detection framework in crowded scenes is proposed based on spatial and temporal contexts. The same authors proposed in (Cong et al., 2013a) a similar approach based on sparse representations over normal bases. Recently, Hu et al. (Hu et al., 2014) proposed a local nearest neighbor distance descriptor to detect anomaly regions in video se-

[1] See `http://www.cs.ubc.ca/~lowe/vision.html` for examples of companies and projects on these topics.

quences. More recently, the authors in (Tran et al., 2014) have proposed a video event detection approach based on spatio-temporal path search. It is also applied for walking and running detection.

In this paper, we adopt a different point of view. We address the question of detecting structural changes or ruptures, which can be seen as a first step for event detection. We propose to use low-level generic primitives and their spatial relationships, and we do not assume a known set of normal situations or behaviors. To our knowledge, the proposed approach is the first one that exploits low-level primitives and spatial relationships in an unsupervised manner to detect ruptures in video. In order to illustrate the interest of spatial relationships, let us consider a car passing another car on a road. For human beings, it is easy to detect and recognize this kind of event. To learn an intelligent system to detect and recognize this event, one solution is to break down this event into the spatial relationships between the objects (the two cars in this case) at many points in time. For example, the car A is behind the car B at the beginning. If the car A wants to pass the car B, the spatial relationships between the two cars rapidly changes from behind state to beside state and then to ahead state. Thus, detecting ruptures in spatial relationships can be important in detecting and recognizing actions or events in video sequences.

We propose to detect in an unsupervised way strong changes (or ruptures) in spatial relationships in video sequences. This rules out supervised learning-based algorithms which require specific training data. This is useful in all situations where an action or an event can be detected based on such changes or ruptures. Here, we use Harris detector (Harris and Stephens, 1988), and/or SIFT detector (Lowe, 2004) to extract low-level primitives, which are suitable to efficiently detect and track moving objects during time in video sequences (Tissainayagam and Suter, 2005; Zhou et al., 2009). In order to associate features points to objects (to compute the fuzzy representation), the algorithm proposed in (Tissainayagam and Suter, 2005; Zhou et al., 2009) can be used. The work presented in this paper is considered as a further analysis step after tracking the objects using feature points. Furthermore, we propose a fuzzy representation of the objects, based on their feature points, to improve the representation of the objects and of the spatial relationships. Then, the structure of the scene is modeled by spatial relationships between different objects using their fuzzy representation. There are several types of spatial relationships: topological relations, metric relations, directional relations, etc. In this paper, we use directional and metric relationships as an example. More specifically, we consider the angle histogram (Miyajima and Ralescu, 1994) for its simplicity and reliability, and similarly the distance histogram. In order to study the evolution of the spatial relationships over time and to detect strong changes in the video sequences, we need to measure the changes in the angle or distance histograms during time. Note that this approach differs from methods based on motion detection and analysis, since it considers structural information and the evolving spatial arrangement of the objects in the observed scene. In the literature, many measures have been proposed to measure the distance between two normalized histograms. Here, we propose to adapt these measures to angle histograms, in order to use them in our method. Finally, a criterion is proposed to detect ruptures in the spatial relationships based on distances between angle or distance histograms over time.

The proposed methods for the fuzzy representation and detection of ruptures in the spatial relationships are described in Section 2. Experimental results are shown in Section 3 in order to evaluate the performance of the proposed approach. Finally, conclusions and future work are presented in Section 4.

2 Rupture detection approach

The proposed approach is divided into two main parts. In the first part, our goal is to estimate a fuzzy representation of the objects exploiting only feature points. In the second one, spatial relationships between objects are investigated, using this representation of the objects. Based on the evolution of the spatial relationships during time, strong changes in video sequences are detected.

The fuzzy representation of the objects using the features points is described in Section 2.1. Specifically, we study the spatial distribution of the feature points that are extracted using a detector such as Harris or SIFT, for a given object. Feature points can be used to isolate and track objects in video sequences (Tissainayagam and Suter, 2005; Zhou et al., 2009). Thus, we suppose that each moving object is represented by a set of interest points isolated from others with the help of such techniques. Here, we propose two different criteria to represent the objects as regions, exploiting only the feature points. The first one is based on the **depth** of the feature points, by assigning a value to each point based on its centrality with respect to the feature points. The second one assigns a value to each point depending on the **density** of its closest feature points. Finally, the depth and density estimations are combined together, to form a fuzzy representation of the object, where the combined value at each pixel represents the membership degree of this pixel to the object. This allows reasoning on the feature points or on the fuzzy regions derived from them, without needing a precise segmentation of the onjects.

In Section 2.2, the computation of the spatial relationships is discussed based on the fuzzy representation of the objects. As an example, we illustrate the concept with the computation of the angle and distance histograms. Then, the existing distances between two normalized histograms are detailed, and the adaptation of these distances to angle histograms is also discussed. Finally, a criterion is defined as the distance between the angle or distance histograms during time, in order to detect ruptures in the spatial relationships.

2.1 Fuzzy object representation

In this section, we detail the estimation of the fuzzy representation based on the feature points.

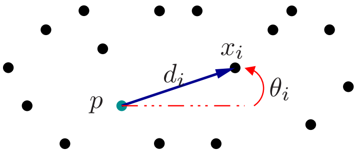


Figure 1: Feature point distribution for a given object.

Feature detection: For a given object, let x_k $(k = 1,2,...,n)$ be the detected feature points. For a given pixel p of the object, let px_i denote the line connecting the pixel p and x_i $(i \in [1...n])$, d_i the distance between p and x_i, and θ_i the angle between $\overrightarrow{px_i}$ and the horizontal line as shown in Fig. 1 $(\theta_i \in [0, 2\pi])$.

Distances d_i and angles θ_i are used to estimate depth and density weights for each object based on the x_i. The depth weight is computed using the angles θ_i, and is denoted by dh. The second weight is computed using the distances d_i, and is denoted by dy. Hereafter, their estimations are described, as well as their fusion.

Depth estimation: In the depth estimation (i.e. centrality), all the feature points are taken into account. Several approaches have been proposed in the literature for depth measures (Hugg et al., 2006), such as simplicial estimation (Liu, 1990), half-space estimation (Tukey, 1975), convex-hull peeling estimation (Eddy, 1982), L1-depth (Vardi and Zhang, 2000), etc. In this paper, we propose a new depth measure which is based on the entropy. For each pixel p, the computed angles θ_i are sorted in ascending order as shown in Fig 2. Let $\tilde{\theta}_i$ $(\tilde{\theta}_j \geq \tilde{\theta}_i$ if $j > i)$ be the sorted angles. We define Δ_i as follows:

$$\Delta_i = \begin{cases} (2\pi + \tilde{\theta}_1) - \tilde{\theta}_n & \text{if } i = 1 \\ \tilde{\theta}_i - \tilde{\theta}_{i-1} & \text{if } i \in [2...n] \end{cases} \quad (1)$$

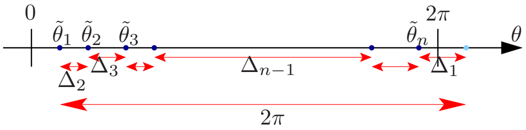


Figure 2: Sorted angles.

Let $p_i = \frac{\Delta_i}{2\pi}$, p_i has two properties: $0 \leq p_i \leq 1$ and $\sum_{i=1}^{n} p_i = 1$. Thus, p_i can be seen as a discrete probability distribution of the angles. Then, the depth weight is defined as the entropy of this probability distribution:

$$\mathrm{dh}(p) = \frac{1}{n} \sum_{i=1}^{n} -p_i \log_2 p_i \quad (2)$$

This depth measure can be explained as follows: let us consider a point q inside the object with feature points distributed equitably around it in terms of directions. In this case, we obtain $p_0 = p_1 = ... = p_n$, and the depth weight of point q is equal to 1 (the highest weight). Otherwise, if the point q is outside the object, the depth weight depends on the angle view (Δ_1 can represent the angle view) and the distribution of the feature points inside the object $(p_2, p_3, ..., p_n)$. If the angle view becomes smaller and smaller (e.g. the point q is moving away from the object), the depth weight of the point q becomes also smaller accordingly.

Fig. 3 shows the representation of several state of the art depth estimations for an object, including our proposal. As we can see, the entropy depth can better represent the shape of the object than the existing depth measures. In terms of computation time, the L1-depth and the

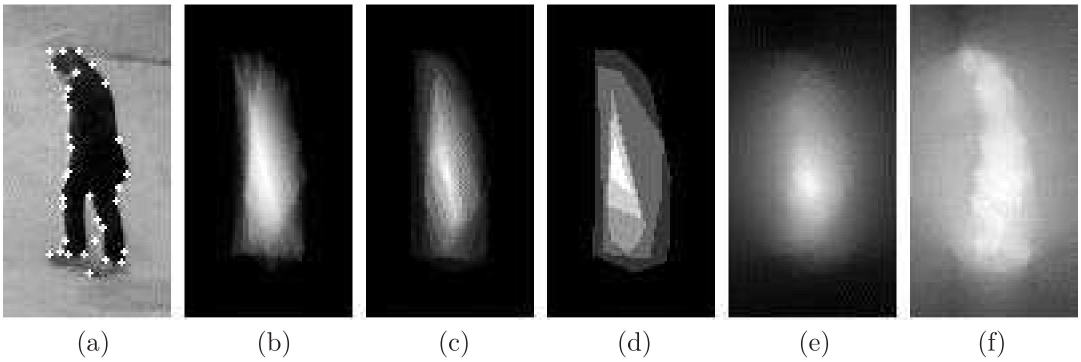


Figure 3: Depth measures: original object with feature points (a), simplicial estimation (Liu, 1990) (b), half-space estimation (Tukey, 1975) (c), convex-hull peeling estimation (Eddy, 1982) (d), L1-depth (Vardi and Zhang, 2000) (e), and the proposed depth (f) (image from PETS 2009 database (PETS, 2009)).

proposed depth are the most efficient ones compared to other measures. Our experimental tests showed that the choice of a particular depth measure has a limited impact on the detection of the rupture. However, the entropy depth measure may present a significant enhancement compared to other depth measures, in the applications that need a precise shape estimation, to describe fine relationships, for example when objects meet.

Density estimation: For density estimation, for a given pixel inside the object, only the neighbor feature points are taken into consideration (feature points within a certain distance r, or k closest feature points). Thus, the distances d_i that are lower than a certain distance r are taken into account to compute the density weight for the pixel p as follows:

$$\mathrm{dy}(p) = \sum_{i=1}^{M} (1 - \frac{d_i}{r}), \text{ where } d_i \leq r \tag{3}$$

where M is the number of points inside the circle of radius r. This radius can be estimated automatically and online, based on statistics on the distances between points, in order to be adapted to the scale of the object. Fig. 4 (c) shows a representation of the density estimation.

Fusion of depth and density estimations: We present a combination approach to fuse the two estimations obtained from depth and density of the feature points. For the sake of optimization, the pixels q that are taken into consideration for the fusion are defined as follows: $\mathrm{dy}(q) > 0$ or $\mathrm{dh}(q) > th$, where th is a given threshold. The obtained estimation of the object is referred to as "fuzzy representation".

Here, the z-score (Carroll and Carroll, 2002) is applied on the two estimations, in order to make them comparable. The z-score is the most commonly used normalization process. It converts all estimations to a common scale with an average of **zero** and a standard deviation of **one**. It is defined as follows: $Z = (X - \overline{M})/(\sigma)$, where $\overline{M}$ and σ represent the average and the standard deviation of the X estimation, respectively. Let Z^{dh} and Z^{dy} be the depth and density estimations respectively, after applying the z-score normalization.

The obtained fuzzy representation, using different fusion operators, are compared with a Ground Truth (GT) where the objects are segmented precisely (see Section 3 for details, and an example in Fig. 4 (g)). The combination approach which gives the best performance consists in using the two operators min and max together as defined in the following expression:

$$\mathrm{F}(p) = \min\left(\max\left(\mathrm{Z}^{\mathrm{dh}}(p), \mathrm{Z}^{\mathrm{dy}}(p)\right), \hat{\sigma}\right) \tag{4}$$

where $\hat{\sigma} = \frac{1}{2\mathrm{th}}$. Then, F is normalized using Min-Max scaling (Han et al., 2006) to obtain the membership function μ_{F} which varies in $[0,1]$. This fusion can be explained as follows: when Z^{dh} (or Z^{dy}) is greater than $\hat{\sigma}$, the membership value $\mu_{\mathrm{F}}(p)$ is equal to 1. Otherwise, $\mu_{\mathrm{F}}(p)$ is less than 1 according to the maximum between them. As an example, Fig. 4 shows different fuzzy representations of the object using min operator, max operator, and Eq. 4 for the fusion. As we can see, the last fusion approach shows the best fuzzy representation of the object according to the ground truth. The obtained fuzzy representations are used to compute the spatial relationships.

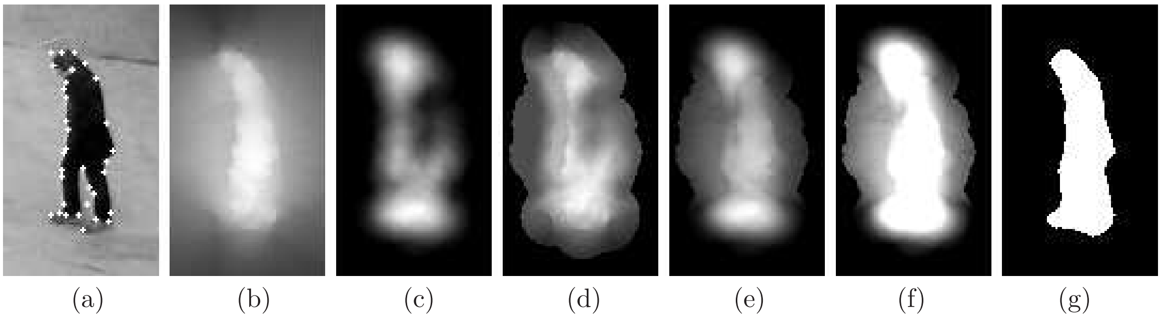


Figure 4: Original object with the feature points (a), depth estimation (b), density estimation (c), fusion using min operator (d), fusion using max operator (e), fusion using Eq. 4 (f), and the object segmented precisely GT (g).

2.2 Spatial relationships and rupture detection

Here, the goal is to estimate the spatial relationships between two objects based on their fuzzy representation. The angle (Miyajima and Ralescu, 1994) and distance histograms are selected as examples to model the spatial relationships. It is important to note that the proposed method also applies to other types of spatial relationships.

Angle histogram: Given two fuzzy regions $A = \{(a_i, \mu_A(a_i)), i = 1, ..., n\}$ and $B = \{(b_j, \mu_B(b_j)), j = 1, ..., m\}$, where a_i and b_j are the elements of A and B, and μ_A and μ_B represent their membership functions respectively, for all possible pairs $\{(a_i, b_j), a_i \in A$ and $b_j \in B\}$, the angle θ_{ij} between a_i and b_j is computed, and a coefficient $\mu_\Theta(\theta_{ij}) = \mu_A(a_i) \times \mu_B(b_j)$ is derived. For a given direction α, all the coefficients of the angles that are equal to α are accumulated as follows:

$$h^\alpha = \sum_{\theta_{ij}=\alpha, i=1,..,n, j=1,..,m} \mu_\Theta(\theta_{ij}) \tag{5}$$

Finally, $h = \{(\alpha, h^\alpha), \alpha \in [0, 2\pi]\}$ is the angle histogram. In our case, the histogram can be seen as an estimate of the probability distribution of the angles. Thus, the obtained histogram is normalized to display frequencies of the existed angles with the total area equaling 1. It is normalized by dividing each value by the sum $R_h = \sum_{\alpha \in [0,2\pi]} h^\alpha$, instead of normalizing by the maximum value (which would correspond to a possibilistic interpretation).

When the objects are represented sparsely by feature points, then $\mu_A(a_i) = 1$ and $\mu_B(b_j) = 1$ (where a_i and b_j represent the feature points on the objects A and B respectively), and the same approach is used to compute the angle histogram between the two sparse objects A and B.

Distance histogram: In this case, all the distances d_{ij} between a_i $(i = 1, ..., n)$ and b_j (j = 1, ..., m) are computed. Based on these distances, the distance histogram is formulated in the same way as the angle histogram:

$$h^l = \sum_{d_{ij}=l, i=1,..,n, j=1,..,m} \mu_L(d_{ij}) \tag{6}$$

where $\mu_L(d_{ij}) = \mu_A(a_i) \times \mu_B(b_j)$ and l represents a given distance value. The obtained histogram is normalized such that the sum of all bins is equal to 1.

Comparison of spatial relationships: There are two main approaches to estimate distances between histograms. The first approach is known as bin-to-bin distances such as L_1 and L_2 norms. The second one is called cross-bin distances; it is more robust and discriminative since it takes the distance on the support of the distributions into account. Note that the bin-to-bin distances may be seen as particular cases of the cross-bin distances. Several distances based on cross-bin distances, such as Quadratic-Form (QF) distance (Hafner et al., 1995), Earth Mover's Distance (EMD) (Rubner et al., 2000), Quadratic-Chi (QC) histogram distance (Pele and Werman, 2010), have been proposed in the literature. We have tested these three distances on different examples, and experiments showed that they were well adapted to angle histograms. Finally, the QF distance was used in our experiments to assess the distance between the angle or distance histograms during time, because of its simplicity. It is defined as follows: $d(h_1, h_2) = \sqrt{ZSZ^T}$, where $Z = h_1 - h_2$ and $S = \{s_{ij}\}$ is the bin-similarity matrix. This distance is commonly used for normalized histograms (the distance histogram for example). Here, we propose an approach to adapt it to the case of angle histograms just by adjusting the elements of the similarity matrix S. We consider that the two histograms h_1 and h_2

defined on $[0, 2\pi]$ consist of k bins B_i. Usually, for a distribution on the real line, the distance between B_i and B_j is defined as follows: $x_{ij} = |B_i - B_j|$, where $1 \leq i \leq k$ and $1 \leq j \leq k$. However, in the case of angle histograms, the distance between B_i and B_j is defined as follows: $x^c_{ij} = \min(x_{ij}, 2\pi - x_{ij})$ to account for the periodicity on $[0, 2\pi]$. Thus, the elements of the matrix S are simply defined, in the case of angle histograms, using x^c_{ij} instead of x_{ij} as follows:

$$s_{ij} = 1 - \frac{x^c_{ij}}{\max_{i,j}(x^c_{ij})} \tag{7}$$

Criterion for rupture detection: Based on the fuzzy representation of the objects exploiting only the feature points, the angle or distance histogram h between two different objects is computed. Let f_i $(i = 0, 1, ..., N-1)$ be the frames of the video sequences, and h_i be the computed angle or distance histogram between the objects A and B in frame f_i. In this paper, we define $y(i) = d(h_i, h_{i+1})$ for each $i = 0, 1, ..., N-1$. This function describes the evolution of the angle or distance histograms over time. If a strong change in the spatial relationships occurs at instant R $(R < N)$, where R denotes the instant of rupture, this means that the angle or distance histogram h_R effectively changes compared to previous angle or distance histograms $(h_i, i < R)$. A rupture is detected according to the following criterion W: $\forall i < R-1$, $y(R-1) - y(i) > t$, and t is a threshold value. Thus, the instant of rupture R can be effectively detected from the analysis the function y.

Here, in order to clearly show the instant of ruptures in the spatial relationships and remove noise, we also show the evolution of the function y filtered by a Gaussian derivative, denoted by g, instead of a simple finite difference. This filter can remove noise and the function g effectively exhibits the instant of the strong changes in the spatial relationships using a threshold approach. This approach is particularly well suited for abrupt changes, leading to clear peaks in the function g, that are then easy to detect (a simple threshold can be sufficient). For slower changes, a multiscale approach can be useful to detect more spread peaks.

3 Experiments and evaluations

To evaluate the performance of the proposed approach, we created some synthetic events (illustrated in Fig. 5, (a) and (b)), and also used a variety of events selected from the PETS 2009 datasets (PETS, 2009) (illustrated in Fig. 5, (c) and (d)). Here, we call "event", some frames that contain a rupture in the spatial behavior. The results of the proposed fuzzy representation are also compared to classical segmentation approaches: a binary segmentation approach (Comanicu and Meer, 2002) and an approach using differences between the background and the actual frame. Then, morphological operations are carried out to remove small objects and fill holes. The last one is used as ground truth (GT) because it produces very precise segmentations.

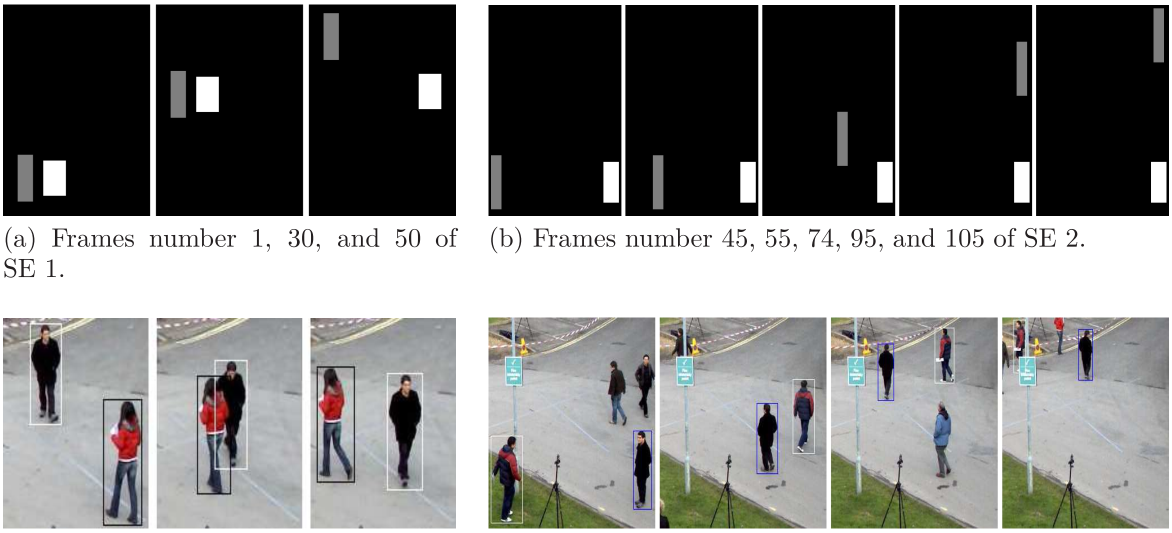
A synthetic event and an event selected from PETS 2006 dataset (PETS, 2006), displayed in Fig. 6, are used to illustrate the proposed approach using the distance histogram.

To associate feature points to objects, here we simply consider the points included in the bounding boxes associated with objects available in the PETS 2009 dataset.

3.1 Parameters tuning

In this section, some results are detailed concerning the tuning of the parameters that are used in the proposed approach. Specifically, we discuss the estimation of the radius r, which is used in the computation of the density estimation. Then, some results are shown for different values of the threshold th, which is used in the combination of depth and density estimations. Finally, we show the effect of the number of bins on the computation of the distance between two angle histograms.

r **parameter:** Fig. 7 shows different estimations of the radius r (normalized) during time. First, all the possible distances d_{ij} among the feature points are computed. The mean, median, and maximum of these distances are computed, as shown in the figure (three first curves). Then, Delaunay triangulation is applied on the feature points, and two other estimations of the radius r are computed, as the mean and median of the lengths of the triangle edges (fourth and fifth curves). Finally, as in (Loménie and Stamon, 2008), the median of all radius of the circumscribed circle around the Delaunay triangles provides the last estimation (last curve). As we can see, the maximum of the distances (third curve) gives the most robust and stable estimation during time. Other experiments on different objects



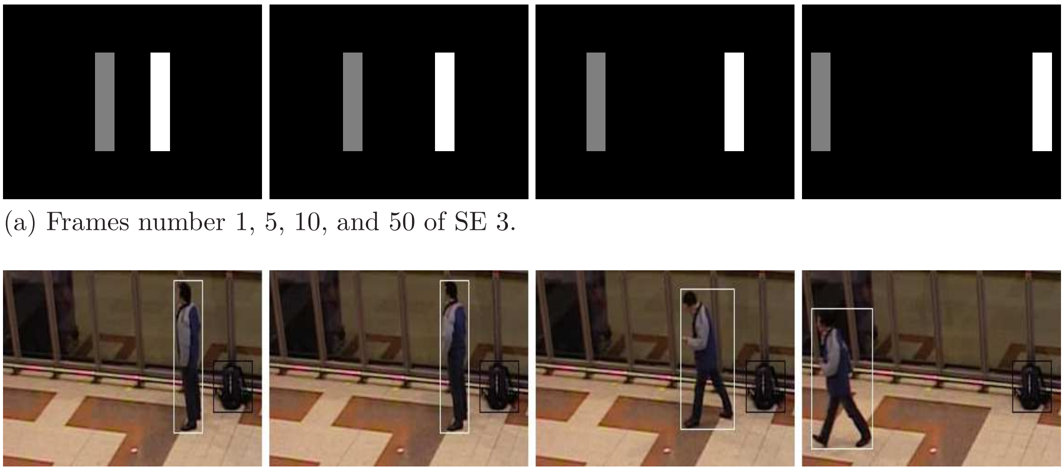
(a) Frames number 1, 30, and 50 of SE 1.

(b) Frames number 45, 55, 74, 95, and 105 of SE 2.

(c) Frames number 450, 462, and 468 of RE 1 selected from PETS 2009.

(d) Frames number 595, 630, 670, and 700 of RE 2 selected from PETS 2009.

Figure 5: Events SE 1 (a), SE 2 (b), RE 1 (c) and RE 2 (d).



(a) Frames number 1, 5, 10, and 50 of SE 3.

(b) Frames number 1955, 2010, 2060, and 2100 of RE 3 selected from PETS 2006 (PETS, 2006).

Figure 6: Events SE 3 (a) and RE 3 (b).

show the same result. Thus, the expression

$$r = \max_{i=1,..,n,j=i,..,m} \frac{d_{ij}}{6} \tag{8}$$

is adopted to estimate the radius r for the density estimation.

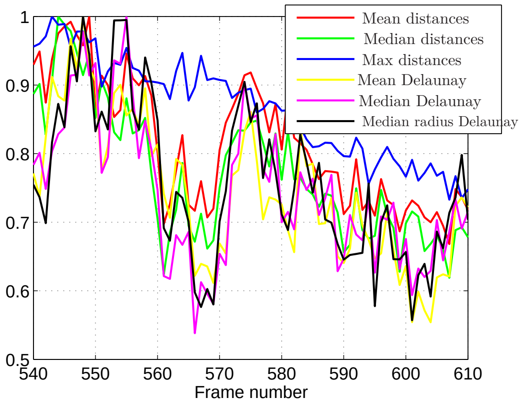


Figure 7: Different estimations of the radius r based on the feature points.

th **parameter:** In the fusion of depth and density estimations, a threshold th is used. Fig. 8 shows the original object with the feature points, the ground truth (GT) of the object, and the fuzzy representation (FR) of the object for different values of th. As we can see, the proposed fusion approach is quite robust to the variation of the used threshold th. In the paper, a value of th equal to 0.5 is used in the combination of depth and density estimations.

Number of bins: In this section, we study the effect of the number of bins (quantification) on the distance between two angle histograms. We defined the function y as the distance between two successive angle histograms in frames f_i and f_{i+1}. Here, we also define $z(i) = d(h_0, h_i)$ for $i = 0, 1, ..., N-1$, i.e. the distance to the histogram in the initial frame, to consider strong changes in the angle histograms. Fig. 9 shows the evolution of the two functions y and z, for numbers of bins of 360, 18, and 6. As we can see, there is almost no difference between 360 and 18 bins, for the two functions. For a number of bins equal to 6, there is a difference compared to 360 and 16 bins for the function z. For the function y, the three curves are almost the same. Thus, the used distance between two angle histograms is robust to the variation of the number of bins.

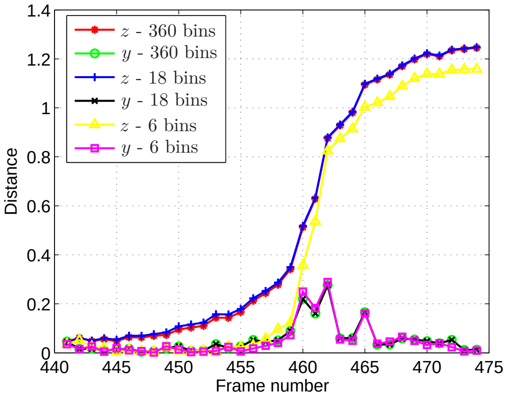


Figure 9: The functions y and z over time using various number of bins.

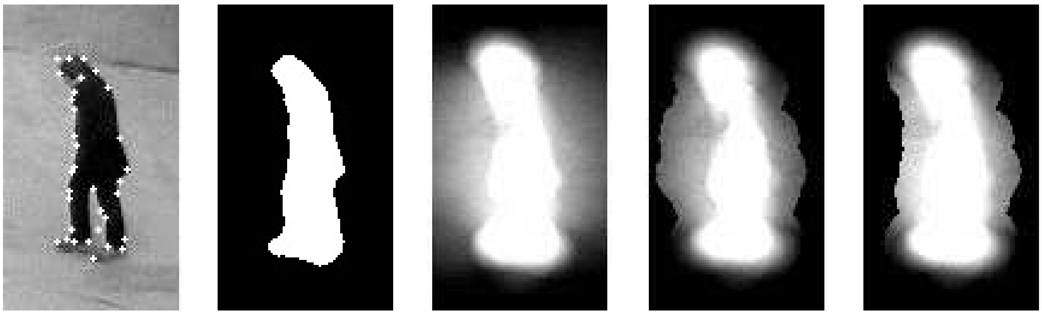
3.2 Ruptures in spatial relationships

We now illustrate how the analysis of the distances between histograms allows us to detect ruptures in spatial relations, both for orientation and distances.

3.2.1 Angle Histogram

Three snapshots of the first synthetic event (SE) are shown in Fig. 5 (a) (two objects moving together and then separately). In this case, there is a rupture in the directional spatial relationships, when the two objects diverge. Fig. 5 (b) shows five snapshots of the second SE. In this event, the object B moves towards the object A (fixed) from the left to the right. Then, the object B changes of direction (frame 74), and when the object B becomes above the object A, it goes towards the top.

Fig. 10 shows the functions y and g during time for the two events SE 1 and 2. For the event SE 1, the function y shows a strong variation at frame number 31. At this instant, there is the rupture in the spatial relationships (the two objects begin to separate). Using the evolution of g over time, the instant of the rupture can be detected by applying a threshold (a threshold of 0.02 can be used to detect the instants of rupture for the SE). For the second SE, we can see two strong variations in the function y; the first strong variation (frame 60) occurs when B changes of direction with respect to A, the second strong variation (frame 90) occurs when B becomes above A and changes its direction towards the top. The function g clearly shows the two strong variations.



Original object GT FR - $th = 0.2$ FR - $th = 0.5$ FR - $th = 0.8$

Figure 8: Original object with the feature points, GT of the object, and fuzzy representations of the object for th equal to 0.2, 0.5, and 0.8 respectively.

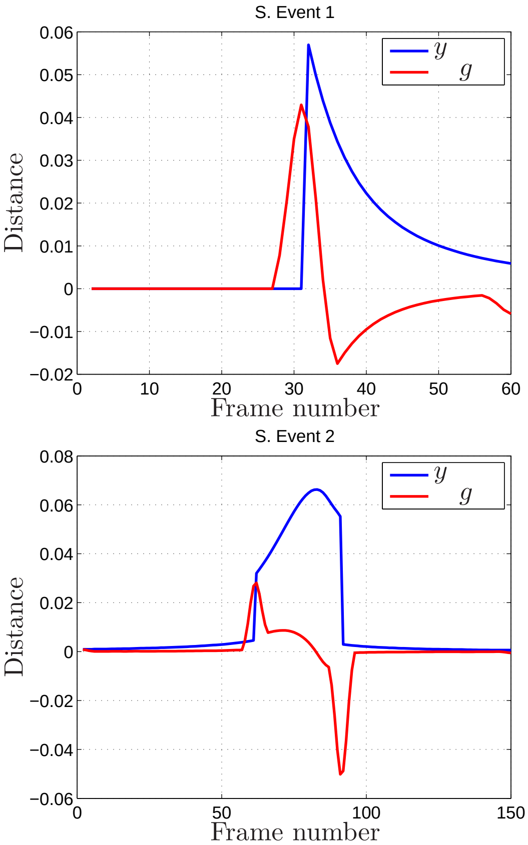


Figure 10: Functions y and g for events SE 1 (top) and 2 (bottom), computed from angle histograms.

Thus, the proposed method can efficiently detect the instants of ruptures in the spatial relationships. Other SE were created and tested using the proposed approach, and similar results were obtained.

Let us now evaluate the proposed detection of ruptures in the spatial relationships in the presence of noise (deformation of objects, etc.) in real events. For the real event (RE) 1 (Fig. 5 (c)), the two persons converge then diverge. Fig. 11 (top) shows the functions y and g over time using the proposed fuzzy representation, for the event RE 1. Two ruptures in the directional spatial re-

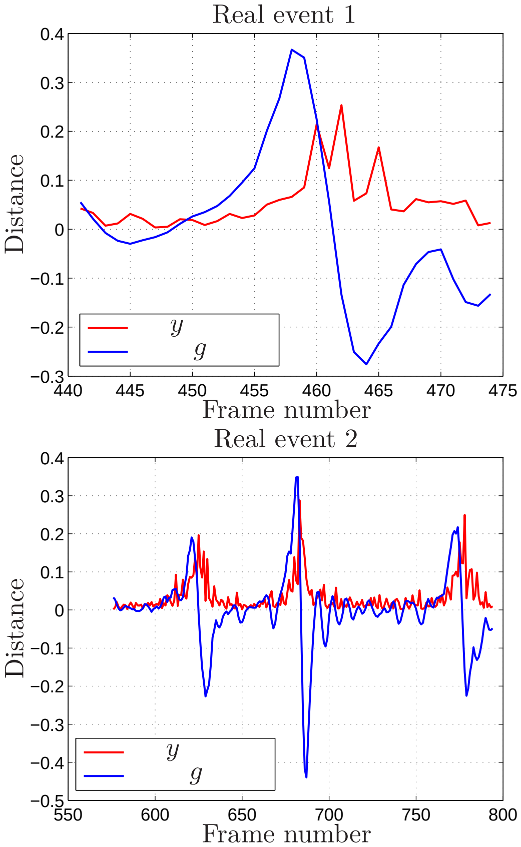


Figure 11: Functions y and g over time, using the proposed fuzzy representation, for the events RE 1 (top) and RE 2 (down), computed from angle histograms.

lationships exist in this event. The first one is when the two persons meet, and the second rupture when the two persons separate. It is clear that the two instants of the ruptures can be efficiently detected using the evolution of g (a threshold of 0.2 can be used to detect the instants of ruptures for the RE). In the event RE 2 (Fig. 5 (d)), the two persons (surrounded by white and blue bounding boxes) converge and diverge several times. In Fig. 11 (bottom), we show the functions y and g over time, using the fuzzy representation of the objects, for the event RE 2. All the ruptures in the directional spatial relationships can be efficiently detected using the function g.

3.2.2 Distance Histogram

Four snapshots of the third synthetic event are shown in Fig. 6 (a). At the beginning of this event, the two objects diverge at a speed of 5 pixels/frame, and at a given instant (precisely at frame 10), the speed of the two objects becomes 10 pixels/frame. Thus, the velocity of the objects is suddenly increased. Fig. 6 (b) shows four snapshots of the third real event selected from PETS 2006. In this event, the luggage is attended to by the owner for a moment, and then the person leaves the place and goes away.

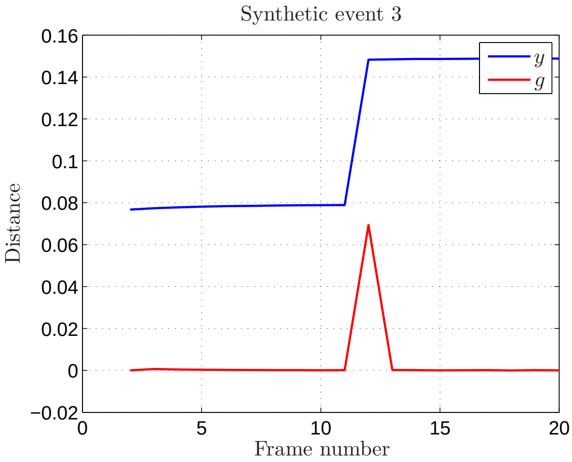


Figure 12: Functions y and g over time, using the proposed fuzzy representation, for the event SE 3, computed from distance histograms.

In Fig. 12, the functions y and g during time for the event SE 3 are shown. As we can see, the function y shows a strong variation at frame number 10, when the velocity of the objects changes. At this instant, a rupture in the metric spatial relationships is detected, using the evolution of g over time.

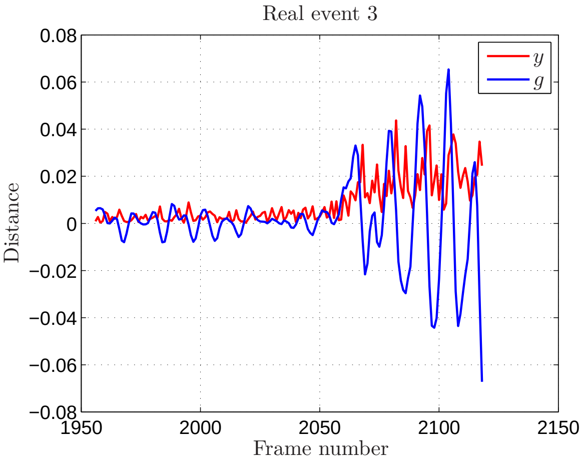


Figure 13: Functions y and g over time, using the proposed fuzzy representation, for the event RE 3, computed from distance histograms.

In the presence of noise, we show in Fig. 13 the functions y and g during time for the third real event. When the person leaves the place and goes away, we can see a strong change in the function y. By nalyzing the obtained results, the instant of rupture in the metric spatial relationships can be detected. These results can be used to indicate events occurring in the video sequences, such as escaping in Fig. 6 (a) and Left-Luggage in Fig. 6 (b).

3.3 Impact of object representation

Here, we show the importance of the fuzzy representation based on a simple feature points representation. Two feature detectors, Harris and SIFT, are tested. Fig. 14 illustrates the function y during time (computed here from angle histograms) for different representations of the objects, for RE 1. The Harris and SIFT features are directly used to estimate the spatial relationships between the two objects and to compute the function y (red and green curves in the figure). In addition, we show in the same figure the evolution of the function y computed on the fuzzy representation of the objects using the Harris and SIFT features (blue and black curves in the figure). As we can see, the evolution of the function y obtained from the fuzzy representation of the objects using the SIFT features (black curve) can significantly reduce the variation of the distance (i.e. less amplitude of the curve) on areas when there is no rupture in the spatial relationships (see Fig. 14, frames 440 to 456) with respect to the SIFT features without computing the fuzzy representation. Thus, the proposed fuzzy representation of the objects before computing the spatial relationships can improve the robustness of the detection of ruptures, based on the observation that SIFT features are more noisy across frames than Harris features in this sequence.

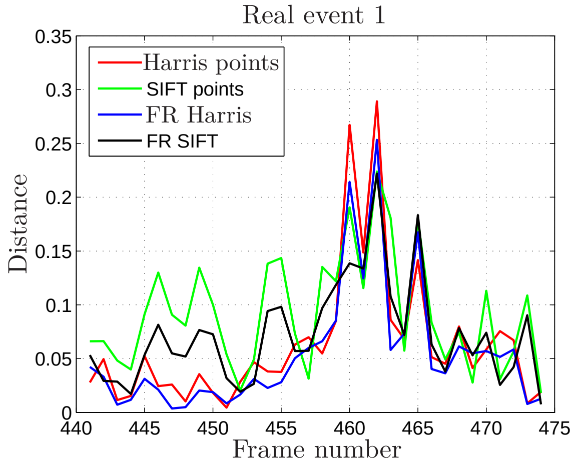


Figure 14: Function y over time, computed from angle histograms, for different estimations of the objects: Harris features, SIFT features, fuzzy representation (FR) of the objects using Harris features (FR Harris) and SIFT features (FR SIFT), for RE 1.

However, noise is present in the function y for all object representations. Assuming that the function y has additive Gaussian noise, the algorithm proposed by Garcia (Garcia, 2010) is used to estimate the variance of the noise of the function y, for the different object representations: Harris features, fuzzy representation of the objects using Harris features (FR Harris), SIFT features, fuzzy representation of the objects using SIFT features (FR SIFT), the binary segmentation using Mean-Shift algorithm (Comanicu and Meer, 2002) and GT.

Tab. 1 shows the variance of the noise in the function y, for the different object representations, for the two events RE 1 and 2. It is clear that the proposed fuzzy representation significantly reduces the variance of the noise, which becomes close to the one of the GT. Especially, for SIFT features, the variance of the noise reduces from 27 to 10 for RE 1, and from 8.9 to 7 for RE 2. In addition, the variance of the noise of the proposed object representation is significantly less than the one of the binary segmentation using Mean-Shift algorithm.

4 Conclusion

In this paper, a new method was proposed to detect strong changes in spatial relationships in video sequences. Specifically, new approaches have been proposed to compute depth and density estimations, based on feature points, as well as fuzzy representations of the objects by combining depth and density estimations. Exploiting the fuzzy representations of the objects, the angle and distance histograms are computed. Then, the distance between the angle or distance histograms is estimated during time. Based on these distances, a criterion is defined in order to detect the significant changes in the spatial relationships. The proposed method shows good performances in detecting ruptures in the spatial relationships for both synthetic and real video sequences.

Future work will focus on further improvement of the proposed method in order to detect other kinds of ruptures, and investigating the use of spatio-temporal relationships. Besides, we will investigate multi-time scale analysis, in order to better detect events that take more time to happen. In addition, proposing a complete event detection framework based on spatial relationships as discriminative features seems to be promising.

ACKNOWLEDGMENTS

This research is part of French ANR project DESCRIBE “Online event detection in video sequences using structural and Bayesian approaches”.

REFERENCES

Advisor (2000). http://www-sop.inria.fr/orion/ADVISOR/. Advisor Project.

Avitrackr (2004). http://www-sop.inria.fr/members/Francois.Bremond/topicsText/avitrackProject.html. Avitrackr Project.

Beware (2007). http://www.eecs.qmul.ac.uk/∼sgg/BEWARE/. Beware Project.

Caretaker (2006). http://www-sop.inria.fr/members/Francois.Bremond/topicsText/caretakerProject.htm. Caretaker Project.

Carroll, S. and Carroll, D. (2002). *Statistics made simple for school leaders: data-driven decision making*. R&L Education.

Comanicu, D. and Meer, P. (2002). Mean shift: A robust approach toward feature space analysis.

Table 1: Estimated variance of the noise ($\times 10^{-4}$) (Garcia, 2010) in the function y, for different object representations, for RE 1 and 2.

Event	Harris	FR Harris	SIFT	FR SIFT	Mean-Shift	GT
RE 1	13	12	27	10	31	12
RE 2	7.7	5.48	8.9	7	31	5.4

IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(5):603–619.

Cong, Y., Yuan, J., and Liu, J. (2013a). Abnormal event detection in crowded scenes using sparse representation. *Pattern Recognition*, 46(7):1851 – 1864.

Cong, Y., Yuan, J., and Tang, Y. (2013b). Video anomaly search in crowded scenes via spatio-temporal motion context. *IEEE Transactions on Information Forensics and Security*, 8(10):1590 – 1599.

Eddy, W. (1982). Convex hull peeling. In *COMPSTAT*, pages 42–47.

Etiseo (2004). http://www-sop.inria.fr/orion/ETISEO/.

Garcia, D. (2010). Robust smoothing of gridded data in one and higher dimensions with missing values. *Computational Statistics and Data Analysis*, 54(4):1167 – 1178.

Hafner, J., Sawhney, H., Equitz, W., Flickner, M., and Niblack, W. (1995). Efficient color histogram indexing for quadratic form distance functions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(7):729 – 736.

Han, J., Kamber, M., and Pei, J. (2006). *Data mining: concepts and techniques*. Morgan Kaufmann.

Harris, C. and Stephens, M. (1988). A combined corner and edge detector. In *Fourth Alvey Vision Conference*, pages 147–151.

Hu, X., Hu, S., Zhang, X., Zhang, H., and Luo, L. (2014). Anomaly detection based on local nearest neighbor distance descriptor in crowded scenes. *The Scientific World Journal*, 2014.

Hugg, J., Rafalin, E., Seyboth, K., and Souvaine, D. (2006). An experimental study of old and new depth measures. In *Workshop on Algorithm Engineering and Experiments (ALENEX)*, pages 51–64.

Icons (2000). http://www.dcs.qmul.ac.uk/research/vision/projects/ICONS/. Icons Project.

Jiang, F., Wu, Y., and K.Katsaggelos, A. (2009). Detecting contextual anomalies of crowd motion in surveillance video. In *16th IEEE International Conference on Image Processing*, pages 1117 – 1120.

Liu, R. (1990). On a notion of data depth based on random simplices. *The Annals of Statistics*, 18(1):405–414.

Loménie, N. and Stamon, G. (2008). Morphological mesh filtering and α-objects. *Pattern Recognition Letters*, 29(10):1571 – 1579.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91 – 110.

Mehran, R., Oyama, A., and Shah, M. (2009). Abnormal crowd behavior detection using social force model. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 935 – 942.

Miyajima, K. and Ralescu, A. (1994). Spatial organization in 2D images. In *Third IEEE Conference on Fuzzy Systems*, pages 100–105.

Pele, O. and Werman, M. (2010). The quadratic-chi histogram distance family. In *European Conference on Computer Vision (ECCV)*, pages 749 – 762.

PETS (2006). http://www.cvg.rdg.ac.uk/PETS2006/data.html.

PETS (2009). http://www.cvg.rdg.ac.uk/PETS2009/a.html.

Piciarelli, C., Micheloni, C., and Foresti, G. (2008). Trajectory-based anomalous event detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(11):1544 – 1554.

Rubner, Y., Tomasi, C., and Guibas, L. (2000). The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121.

Saleemi, I., Shafique, K., and Shah, M. (2009). Probabilistic modeling of scene dynamics for applications in visual surveillance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(8):1472 – 1485.

Tissainayagam, P. and Suter, D. (2005). Object tracking in image sequences using point features. *Pattern Recognition*, 38(1):105 – 113.

Tran, D., Yuan, J., and Forsyth, D. (2014). Video event detection: From subvolume localization to spatio-temporal path search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(2):404 – 416.

Tukey, J. W. (1975). Mathematics and the picturing of data. In *International Congress of Mathematicians*, volume 2, pages 523–531.

Vardi, Y. and Zhang, C.-H. (2000). The multivariate l1-median and associated data depth. *National Academy of Sciences*, 97(4):1423–1426.

Visam (1997). http://www.cs.cmu.edu/∼vsam/. Visam Project.

Zhou, H., Yuan, Y., and Shi, C. (2009). Object tracking using SIFT features and mean shift. *Computer Vision and Image Understanding*, 113(3):345 – 352.