Ontology-Based Trajectory Analysis for Semantic Event Detection

Alexia Briassouli, Stamatia Dasiopoulou, Ioannis Kompatsiaris
Informatics and Telematics Institute
Centre for Research and Technology Hellas
abria@iti.gr,dasiop@iti.gr,ikom@iti.gr

Abstract

The extraction of human centered descriptions, matching end users cognition, and specifically the detection and identification of events in videos is a particularly challenging problem, due to the volume and diversity of both the automatically extracted low-level features and the corresponding high-level information conveyed. Numerous efforts have begun, attempting to bridge the semantic gap between low-level data and higher level descriptions, often resorting to domain-specific learning-based approaches. In this paper we present a novel, generally applicable approach, for hierarchical semantic analysis of spatiotemporal video features (trajectories) in order to localize and detect events of interest. Dynamically changing trajectories are extracted by processing the optical flow, based on its statistics. The temporal evolution of the trajectories' geometrical and spatiotemporal characteristics forms the basis on which event detection is performed. This is based on the exploitation of prior knowledge, which provides the formal conceptualization needed to enable the automatic inference of high level event descriptions. Experimental results with a variety of surveillance videos are presented to exemplify the usability and effectiveness of the proposed system.

1. Introduction

The advances of signal processing, networking and hardware have led to the widespread use of digital multimedia applications, not only in research or academic environments, but in all aspects of peoples' daily lives. This has led to an increased interest in higher level, content-based approaches to the manipulation, processing, access and dissemination of digital data. Thus, the semantic analysis of digital imagery, video, audio has attracted significant attention. The processing of digital multimedia is focused on extracting descriptions related to the end users' perception of events, objects. However, this is a very challenging problem, due to the sheer volume and diversity of digital documentation that is available today, and to the variety of high-level information conveyed in it. There have been efforts to address the problem of "bridging the semantic gap" between low-level features and higher level meanings, but they often resort to domain-specific learning approaches, which is necessary, due to the diversity of the available information. Nevertheless, efforts are under way to develop more generally applicable methods, that can be applied with small modifications to different cases of digital documentation.

This paper presents a novel method for the spatiotemporal semantic analysis of low-level video features, in order to localize and detect events of interest. Ontologies are designed to allow the extraction of meaningful semantics from the generic, machine-level processing of digital multimedia data. Specifically, dynamically changing trajectories are extracted from the processing of the estimated optical flow's statistical characteristics. Trajectory information is accumulated and the beginning or ending of events is detected via changes in its spatiotemporal distribution. The geometrical characteristics of the accumulated trajectory information are related to concepts in the knowledge structures (ontologies) and thus high level information can be extracted from the low-level processing. The data is processed in a hierarchical manner, so that the various processing stages are more generic and their combined use leads to a system tailored to specific event detection. Thus, knowledge and consequently derived descriptions, propagate along successively higher abstraction levels.

2. Ontology Design for Event Detection

The low-level features extracted from the frames, such as object motion (flow fields), color and texture features, hold significant information about the analysed content, but cannot be directly used to identify the conveyed semantics, e.g., the events taking place. For this reason, we design ontologies [9], whose goal is to represent and capture the extracted video data in a manner that is as generally applicable as possible, and at the same time useful for specific applications.

This leads us to design modular ontologies that combine the temporal characteristics of the motions taking place, and their spatial description. The appropriate combination of different ontology modules leads to conclusions about the type of event taking place. In the following we discuss the kind of descriptive knowledge that can be extracted from video and how it relates to higher level, domain specific semantics. These observations form the basis for the design of the employed ontologies.

2.1. Activity Areas, Trajectories and Events

The video processing stage first leads to the determination of the frames at which events begin and end. This leads to the extraction of the subsequence of interest to be processed. The combination of this information with the appearance of areas of activity, motion information, and the respective trajectories is mapped into knowledge structures, leading to the detection of specific events.

The processing of the velocity vectors estimated via optical flow leads to activity areas 3.3, which are binary masks of the pixels that undergo a motion in the subsequence being processed. The shape of these masks can be very useful for the determination of the events taking place. In surveillance videos, there are many cases of people walking across the area being examined, so the processing of their velocities leads to trajectories. Similarly, in other videos, such as sports videos (tennis, ping pong, soccer), the balls being hit lead to similar, linear or curved "narrow" trajectories. On the other hand, areas of lateral player motion, in the case of tennis for example, lead to activity areas with a different, larger shape. In the case of surveillance videos, when people stop moving towards a specific direction, but continue moving locally, e.g. if they have stopped to walk, fight, leave a package, fall down, there will be a larger, non-linear activity area. Note that by "linear" we refer to an area of activity that originates from a trajectory, and it can be strictly linear, or a curved line.

These two shapes of activity areas, combined with the time instants at which they appear, can immediately lead to certain conclusions about the events taking place. If in the tennis domain there is an activity area shaped like a linear trajectory followed by a larger area, corresponding to lateral motion, and it is followed by another linear trajectory, the ontology maps these events to a successful return of the tennis ball. The same can be applied for any sport involving hitting a ball. In the case of surveillance, the same time sequence of activity areas leads to the conclusion that during the central frame subsequence an event took place. This event may be benign (people stopping to talk), or not, e.g. it may indicate a fight. Then, further processing of the central area of activity takes place, in order to arrive at more specific conclusions. If during these frames there is significant motion (i.e. if the velocity vectors have high magnitudes), the ontology maps this into a "high activity event", which indicates more danger than a "low activity event". Indeed, in the experiments 4.3 we extract such an area, where the velocities during a fight are higher.

2.2. Ontology Deployment

Based on the aforementioned, we followed a modular architecture and developed an ontology infrastructure that couples the different types of knowledge encoded, namely activity areas' characteristics (focus is on trajectory related features), spatiotemporal, and domain-specific information. We stress that the main goal is the development of reusable ontologies that can be effectively applied in different applications.

Regarding the trajectory related knowledge, the developed ontology has a rather simple, yet well-defined, structure that can be easily extended to accommodate more elaborate descriptions. For example, in the case of typhoon movement studies, additional information about the trajectory curve characteristic, such as variation rate, etc. would be required, to obtain accurate behavioral descriptions. As illustrated in Fig. 1, the main classes are the *ActivityArea*, *Trajectory*, *Direction* and *Motion*. Since in our current investigation, shape information is addressed in a quite coarse granularity, the *Shape* class has not been further specialized. The same holds for the other subclases of the so called *StaticFeature* class, which encompasses visual characteristics. In an application, where color and texture segmentation and descriptors would be utilized, the extension of the ontology with respective definitions would allow, additional functionalities, such as object detection that would enrich the contextual knowledge leading to more complete domain-specific descriptions.

The trajectory ontology constitutes a very raw, in terms of end user cognition, semantic conceptualization. However, it allows to map the extracted video features into machine processable descriptions and make their semantics explicit. Following a similar rationale, an ontology module has been developed to account for spatiotemporal aspects of knowledge. Spatial and temporal information is of particular significance in video semantics extraction, as the spatial and temporal dependencies and associations determine largely the conveyed meaning. In the application domains that we currently handle, i.e., tennis and surveillance video, the temporal knowledge that is of particular interest is the time sequence of the different actions, i.e., relations like *precedes*, *overlaps* etc. Regarding spatial relations, current focus is on topological aspects, e.g., *touches* and *distinct*, that allow the identification of events such as "people meeting, standing for a while, and then walk away apart". Direc-

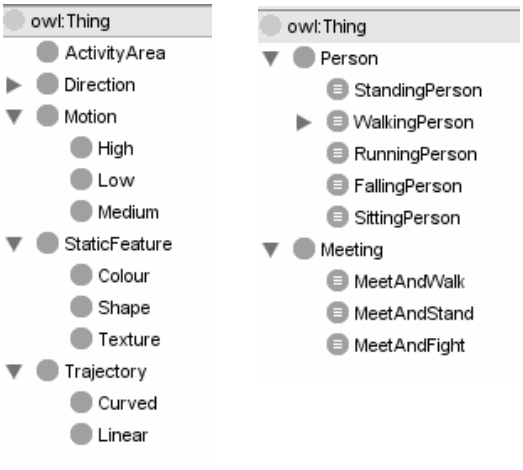

Figure 1. Top level concept hierarchy.

tional relations could provide additional information for extracting more detailed video annotations, like "two people meeting, the second was approaching the one from his left", which could be of interest for a number of applications, including surveillance. As such, both topological and directional relations have been defined, although the exploitation of directional ones is rather limited currently. We chose to represent both types of relations as ontology properties, as this matches better the intuitive semantics and avoids the definition of rigid class definitions structures. This choice, as will be detailed in the sequel affects as well the inference capabilities required, and necessitates the use of rules as well, on top of the ontology definitions.

Having available the conceptualizations to precisely define low-level and spatiotemporal knowledge, what is needed are the domain specific semantic associations that translate combinations and sequences of low-level and spatiotemporal descriptions into meaningful domain interpretations. To accomplish this we need to define domain specific ontologies, that provide more than the vocabulary to describe salient the objects and relations. More specifically, an interpretation oriented approach to ontology engineering needs to be adopted. Taking for example the surveillance video domain, the important events relate to activities such as walking, standing still, running or sitting, people meeting, the ending of such meetings, i.e., if the walk away together or split again, and so on. As such we first define in the ontology all appropriate concepts and properties to model such events (Fig. 1). In addition, axioms and rules have been added to provide the necessary linking with the low level representations though spatiotemporal relations. For instance, the *StandingPerson* class is defined as an activity area having minor motion, while the *WalkingPerson* class similarly but with medium motion values. We note that the exact semantics of low and medium motion, are domain dependent, i.e., become concrete per domain. With the above definitions available, the *FallingPerson* class is defined as the temporal subsequence of a walking or standing person by an instance of an area undergone vertical movement. To handle this kind of *triangle* relations, i.e., relations between individuals that are connected through a common individual, rules need to be introduced.

Following such modular, ontology-based representation, allows to apply the same video analysis approach to different domains, requiring only for the provision of the respective domain semantics, and not tedious re-training and re-adjustment of parameters. It must be noted, that to avoid issues raised by existing reasoners with respect to concrete domains handling, the instantiation of the temporal relations that hold among the produced activity areas individuals, is performed by analysis in the same way it populates low level trajectory related concepts and properties.

3. Video Processing

A very characteristic low-level feature of surveillance videos that can lead to useful conclusions about the events possibly taking place in them is the motion detected in the video. The proposed method is based on the distribution of the optical flow between pairs of frames in a sequence. When a pixel undergoes a displacement, there is a change in its luminance, which is detected by optical flow estimation methods. These methods are based on the assumption of constant luminance, and as a consequence are sensitive to illumination changes in a video sequence. Thus, there will be optical flow estimates on pixels that have not actually undergone motion, but whose luminance changes because of camera measurement noise, or other sources of illumination variation. Naturally, there has been significant work on the application of smoothness constraints on the flow estimates, to ensure robustness against small illumination variations [4], [3]. In this work, the optical flow is estimated using pyramidal techniques, a coarse-to-fine version of the Lukas-Kanade algorithm is applied, in order to extract the flow with accuracy, even for large displacements.

Our approach is based on the realistic assumption that the non-zero optical flow estimates are caused by measurement noise, which is approximated by a Gaussian distribution [2]. The distribution of each pixel's flow is used for two purposes. Its first use is for detecting frames that are considered as "candidates" for the beginning of an event. This is determined by the statistical processing of the flow estimates accumulated over a subsequence of frames, as detailed in Section 3.1. The spatial distribution of pixels that undergo motion over each subsequence is determined by the extraction of "activitiy areas", a process that is detailed in

Section 3.3.

3.1. Temporal Event Localization

As explained in the previous section, the optical flow estimated over a video sequence is often noisy, due to varying luminance values. Thus, the inter-frame flow estimates at each pixel can originate from that pixel's velocity, and noise, or only from noise. This leads to the following two cases for the distribution of the flow of each pixel $\bar{r}$:

$$\begin{aligned} H_0 &: \bar{r} \sim f_{static}(\bar{r}) \\ H_1 &: \bar{r} \sim f_{active}(\bar{r}), \end{aligned} \tag{1}$$

where $f_{static}(\bar{r})$ is the distribution over time of pixels that do not actually undergo motion, and $\sim f_{active}(\bar{r})$ is the distribution of "active" pixels.

In practice, the noise sources during the imaging process are unknown, so f_{static} is unknown a priori. Similarly, the diversity of the possible motions a pixel can undergo over a subsequence, even in the same video, does not allow the prior modeling of f_{active}. Nevertheless, it is realistic to assume that the flow estimates caused by illumination variations have a lower variance than the actual optical flow that is caused by pixel motion. The changes in the luminance of pixels that undergo motion are going to be higher, on average, than the luminance changes caused only by noise. Larger illumination variations will create outliers in the flow values (including the noisy flow values), which appear as heavier tails in the resulting data distributions.

This was verified experimentally as well. The empirical distribution for a pixel that did not undergo any motion was compared against the distribution of a moving pixel over the video sequence. The resulting histograms, shown in Fig. 2 on a logarithmic scale, demonstrate that, indeed, the flow for active pixels has heavier tails than the flow for static pixels. This essentially means that the optical flow for active pixels will contain more significant variations than the noisy flow estimates of pixels that are actually static. It should be noted that Fig. 2 has been produced by first estimating the area of active pixels in a video, using the method detailed in the section that follows (Sec. 3.3). The flow for all pixels in the activity area, over all video frames, is extracted. The empirical "temporal distribution" of each pixel's flow (i.e. the way its values vary during the video sequence) is estimated, and the average of all pixels' flow distributions is evaluated. This is considered to be the empirical approximation to f_{active}. Similarly, the temporal flow distribution for the background pixels is extracted, and averaged, in order to estimate the average static pixel flow distribution, i.e. f_{static}.

Empirical Likelihood Sequential Hypothesis Testing

The heavy tailed nature of the activity area's pixels' distribution is used in order to detect the time instants (frames)

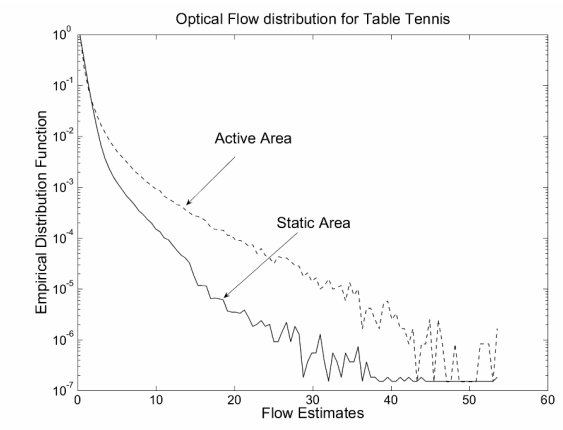

Figure 2. Optical Flow distribution for static and active areas.

at which an event begins. Initially the activity and static areas are separated, as described in Section 3.3, using all frames of the video sequence. This procedure separates active and static pixels, but does not indicate at which time instants each activity begins. For this purpose, we first accumulate ten video frames and estimate the empirical optical flow distributions for the activity and static areas. Afterwards, video frames are sequentially processed, i.e. their flow is estimated, and this new estimate is matched against the previously estimated probability distributions. In the experiments, the frame counting begins after frame 10, as the actual event time-detection begins then.

Since f_{static} and f_{active} are empirically estimated, they are non-parametric (they are not considered to be modeled by a known parameter-dependent distribution, in order to ensure the generality of the proposed method). Thus, we determine if each new random variable (optical flow estimate) belongs to a different distribution than previously, based on the experimentally determined probability distributions for the two frame subsequences being examined. The solution to this problem leads to the determination of the frames at which an event begins or ends, and can be addressed using empirical likelihood tests [5], [6]. The difference between the experimentally evaluated distributions can be detected based on Kolmogorov-Smirnov testing [1], where the "known" distribution is considered to be the empirical distribution of the previous subsequence.

If the pixel was in a static area in the previous subsequence, and it is know assigned to an activity area, the frame being examined is a candidate for the start of an event. In order to ensure the robustness of the proposed method, we assign a frame as the beginning of an event if this new assign-

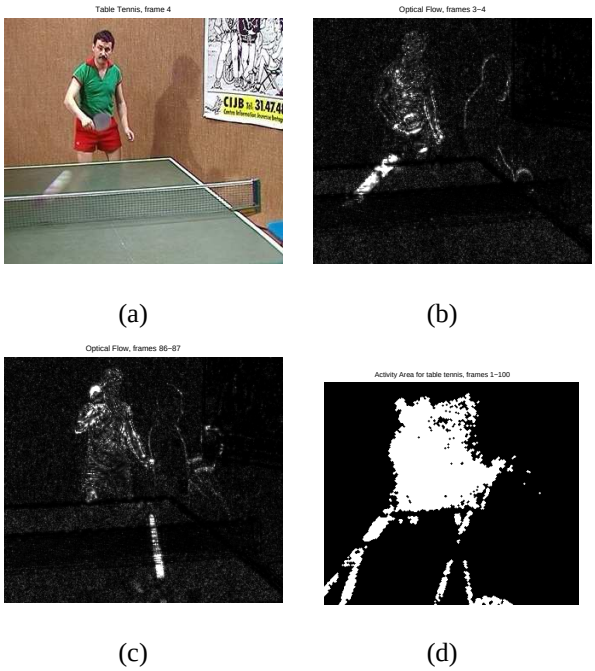

Figure 3. Table Tennis. (a) Frame 4. Optical flow between frames: (b) $3-4$**. (c)** $86-87$**. (d) Activity Area for all frames.**

ment remains valid over five frames. Experiments showed that this is a reasonable assumption. Similarly, if a pixel from an activity area is matched to a static in the frame being examined, this indicates that the activity in that pixel has stopped. Thus, the spatiotemporal localization of events is achieved, since the frames at which activities begin and end are estimated, simultaneously with the pixels at which they occur.

3.2. Spatial Event Localization

In the previous section we described the method for determining the temporal locations of events, i.e. at which frames an event begins and at which frames it ends. In order to find the spatial location of the moving objects in the video sequence, we estimate the optical flow between pairs of frames, using the Lukas Kanade algorithm. Since it is based on the constant illumination assumption, optical flow suffers from inaccuracies introduced by illumination changes that are not introduced by object motions (e.g. lighting changes, measurement noise). Although the Lukas Kanade method is more robust to these inaccuracies than other methods, the flow estimates are still noisy. Also, their values are higher near motion boundaries, and negligible in smooth areas of moving objects. For a video of a person playing table tennis, shown in Fig. 3(a), the flow estimates between two characteristic pairs of frames, shown in Fig. 3(b)-(c), indeed contain significant values only at the moving object boundaries.

We take advantage of the velocity estimates' noise, to extract binary activity masks in each video sequence, with the pixels that undergo displacements during several (if not all) frames. Since we have many samples of this noise (it affects the flow estimates over all frame pixels, over many frames), we approximate it by a Gaussian distribution. Thus, finding moving pixels is reduced to testing if the accumulated velocity estimates follow a Gaussian distribution. For a random variable y, the classical measure of non-gaussianity is the estimation of its fourth order cumulant, also known as the kurtosis: $kurt(y) = E\{y^4\} - 3(E\{y^2\})^2$. The fourth order moment of Gaussian random variables is given by $E\{y^4\} = 3(E\{y^2\})^2$, so ideally the kurtosis of a Gaussian random variable should be equal to zero. Motivated by this, we accumulate the flow estimates v for each pixel, over the frames of the subsequence under examination, and characterize each pixel according to:

$$\begin{cases} \bar{r} \in action\ area\ if\ E\{v^4\} = 3(E\{v^2\})^2 \\ \bar{r} \in background\ if\ E\{v^4\} \neq 3(E\{v^2\})^2. \end{cases} \quad (2)$$

We then estimate the kurtosis of each pixel's flow estimates over the frames being examined. Since the Gaussian model is only an approximation, we do not expect the kurtosis to be zero, but we do expect it to be significantly higher at pixels that have undergone motion. We consider that pixels whose flow has kurtosis above 10% of the mean kurtosis have been displaced. These pixels form an "activity area". In Fig. 3(d) we show the activity area extracted by processing all the video frames of the tennis sequence. Obviously, the activity area not only correctly localizes the pixels that undergo motion, but it also has a shape that is very characteristic of the event taking place. Consequently, it can be used in combination with the ontology designed in Sec. 2, for the extraction of semantics, such as event detection and characterization, from the video sequence.

3.3. Activity Area Shape Extraction

As explained above, the activity areas for the subsequences of interest in the video contain shapes and curves, that are indicative of the event taking place, and can be used in combination with appropriately designed knowledge structures (Sec. 2), for semantic event detection. In order to describe the shapes of the activity area in a manner compliant with the requirements of currently used standards, namely MPEG-7 [8], we use a Region Shape Descriptor [7] that consists of 35 quantized coefficients of the area's Angular-Radial Transform (ART). This descriptor is

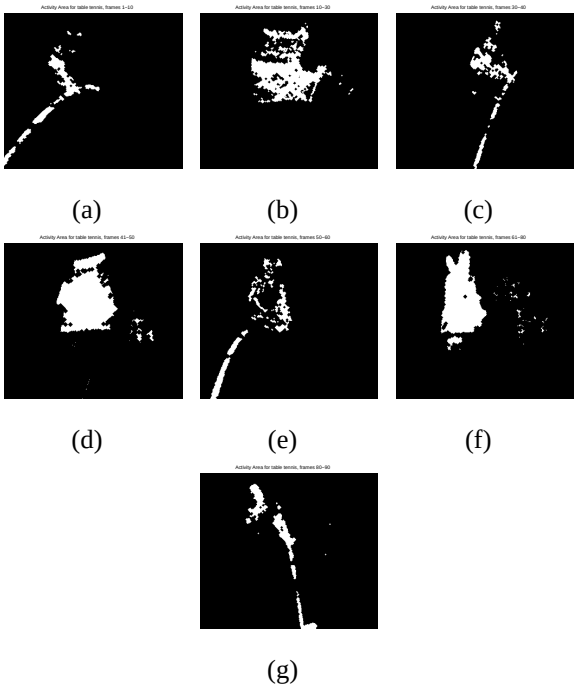

Figure 4. Activity Areas for Table Tennis with player motion with and without ball trajectories.

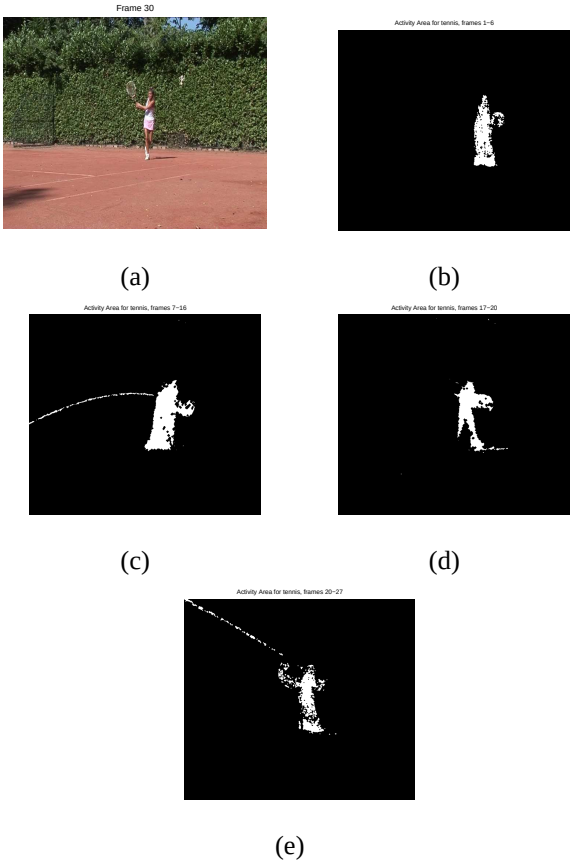

Figure 5. Tennis. (a) Frame 4. Optical flow between frames: (b) $3-4$**. (c)** $86-87$**. (d) Activity Area for all frames.**

particularly well suited for the shapes of the extracted binary activity areas, as it allows the shape of the object to be a single region, a set of regions, or even contain holes. The ART descriptor also has the advantages of being invariant to rotation, scale and position, and is generally robust to noise along an object's contour. The resulting feature vectors for each region is a number of normalized coefficients, and the comparison between such descriptors is simply uses the L_1 distance between them [10].

In the experiments, we estimate the mean absolute L_1 distances between the shape descriptors of areas that correspond to similar and to different activity areas and, consequently, events. The experimental results verify that, indeed, these descriptors allow us to discriminate between different regions of activity and, thus, classify detected events.

4. Experiments

Experiments were conducted with different real video sequences, where events of interest take place. The sequences concern a variety of contexts, such as surveillance, sports, traffic, but the same method is applied to all of them, successfully extracting the relevant event features.

4.1. Table Tennis Trajectory Analysis

Experiments were conducted with a video of a person playing table tennis. The activities of interest are determined by the motions of the ball and the player. The times (frames) at which events begin and end are extracted by comparing the empirically estimated probability densities for frame subsequences. This leads to a temporal segmentation of the video sequence, which we have divided into subsequences of interest, i.e. sets of frames during which different activity occurs in different frame pixels. The frames separating the subsequences of interest are frames 10, 30, 40, 50, 60, 80.

Binary masks of the activity areas corresponding to each time segment between these frames are then extracted. Their shapes are characteristic of the activity taking place in the video, as seen in Fig. 4. The activity area corresponding to a ball's trajectory is a curved line, and is often extracted along with a mask corresponding to the player's silhouette.

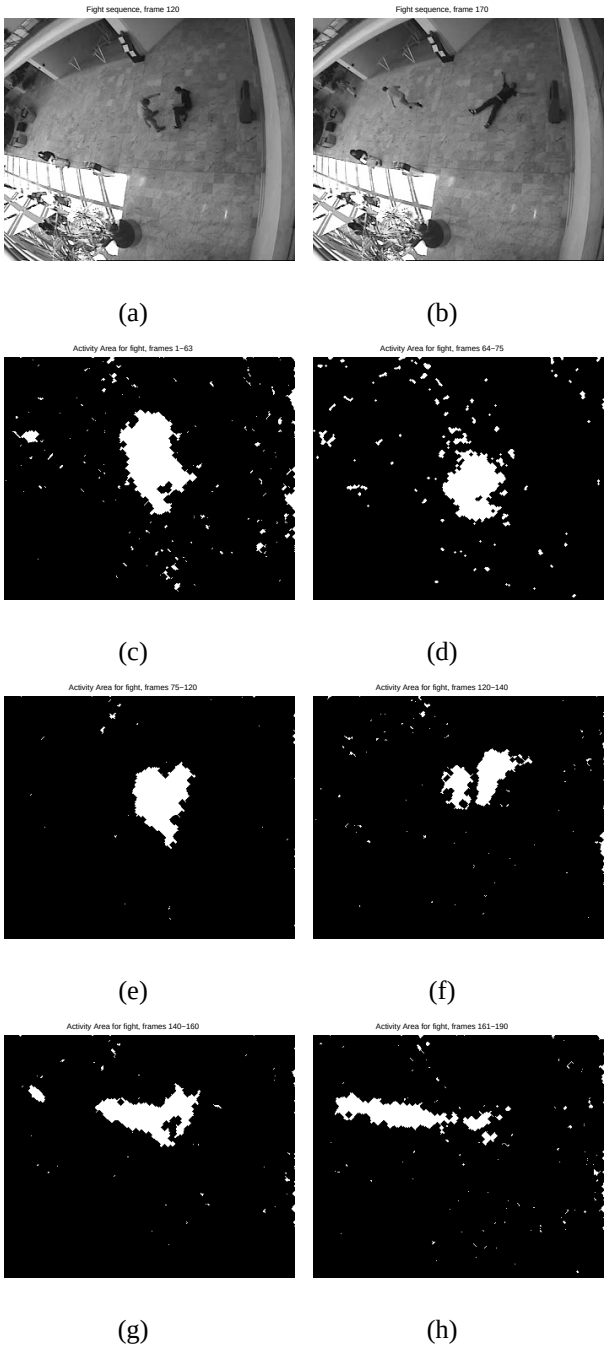

Figure 6. Fight sequence. (a) Frame 120**. (b) Frame** 170**. (c) - (h) Activity Areas.**

This is because the player is moving as the ball is arriving, or right after he has hit it. In this case there is only one subsequences where there is no ball motion, but only the player moving laterally, shown in Fig. 4(b). The activity area from those frames is larger activity area, and does not contain any curved trajectories. The shape characteristics of the activity areas are extracted following the region shape extractor of the MPEG-7 standard, as a vector of 35 normalized ART coefficients.

For sake of comparison, we estimate the absolute L_1 distance between the shape descriptor of the first activity area (a trajectory) with the other five activity areas. The resulting distances are given by:

$$[0.5714,\ 0.2286,\ 0.4823,\ 0.0011,\ 0.5532,\ 0.1429].$$

We see that, indeed, the descriptors for the activity areas of Fig. 4(b), (d), (f) are quite different from those corresponding to ball trajectories. Consequently, they can be considered as a reliable feature for the discrimination between ball trajectories and player lateral motion, and used to extract higher level information, to be used as input in the knowledge structure.

4.2. Tennis Trajectory Analysis

This experiment concerns a similar video, with a girl playing tennis. As before, the frames between which events are expected to take place are detected, and in this case are frames 6 and 20. Fig. 5(a) shows a frame of the video under examination and Figs. 5(b)-(d) the activity areas. As before, there are activity areas corresponding to the ball's trajectory and some player motion, and an area corresponding only to the player's lateral movement. The shape descriptors are extracted and their comparison leads to the following absolute mean L_1 differences between the shape descriptors of the first activity area, and the other three:

$$[0.1143,\ 0.0571,\ 0.2000].$$

As expected, the activity area of Fig. 5(c), where there is no ball trajectory, is the closest to the first activity area.

4.3. Surveillance Event Detection: People Walking and Fighting

Event detection is particularly important in surveillance applications, as the behavior of individuals walking, running, stopping, meeting may be indicative of suspicious activity. For this reason, we analyzed a surveillance video, where two people are fighting, and when one of them falls down, the other runs away (Fig. 6(a), (b)). It should be noted that these are benchmark videos, taken from the PETS-CAVIAR dataset. The temporal event localization

is achieved using flow statistics over the video sequence (Sec. 3.1). The sequence is divided into segments by frames 63, 75, 120, 140 and 160. The resulting activity areas are shown in Fig. 6(c) - (h). The first 160 frames correspond to the case where the people are fighting. Although these activity areas have a similar shape, they occur at different spatial locations in the video frames. Since the time instants at which events begin and end are detected based on the change in the pixels' statistical distributions (regardless of their spatial location), when the fight takes place in different areas, it corresponds to different events. After frame 161, one person starts running away (so again, different pixels are "activated"), leading to an activity area with a different shape. We extract the six MPEG-7 region shape descriptors for these six regions and find their absolute mean L_1 distance between the area of Fig. 6(a) with the areas of Figs. 6(b)-(h):

$$[0.0571,\ 0.1714,\ 0.482,\ 0.1712,\ 0.8]$$

. Indeed, the activity areas of Fig. 6(b), (c), (e) have a small difference from that of Fig. 6(a), where the two people are fighting. The activity area of Fig. 6(d), where one person falls has a larger difference from the first activity area, since it consists of two different active regions (introduced by the motions of one person leaving and one person falling). Finally, the activity area corresponding to the person running away has the highest distance, as its shape is the most different.

Once the activity areas are extracted and compared, as described above, they can be used in combination with the video frames in order to localize the events of interest. Thus, we mask video frames corresponding to the beginning or end of events, using the respective binary activity areas. As Fig. 7 shows, this leads to the correct localization of events of interest, namely the fight, one person falling, the other person waving, and the other person running away.

5. Conclusions

This paper presents a novel method for the spatiotemporal semantic analysis of low-level video features. Ontologies are utilised to allow the extraction of meaningful semantics from the generic, machine-level processing of digital multimedia data. Specifically, dynamically changing trajectories are extracted from the processing of the estimated optical flow's statistical characteristics. The geometrical characteristics of the accumulated trajectory information are related to concepts in the knowledge structures, which provide the formal conceptualization needed to enable the automatic inference of high level descriptions. The data is processed in a hierarchical manner, so that the various processing stages are more generic and their combined use leads to a system tailored to specific event detection.

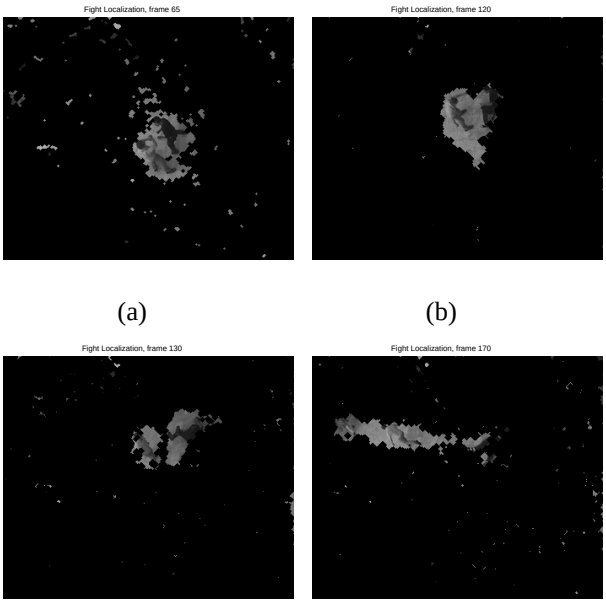

Figure 7. Fight event localization.

References

[1] Chakravarti, Laha, and Roy. *Handbook of Methods of Applied Statistics*. John Wiley and Sons, Volume I, Boca Raton.

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

[3] B. Horn and B. Schunck. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.

[4] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of Imaging understanding workshop*, pages 121–130, 1981.

[5] A. Owen. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, (75):237–249, 1988.

[6] A. Owen. *Empirical Likelihood*. Chapman and Hall/CRC, Boca Raton., 2001.

[7] M. P. Salembier and T. Sikora. *Introduction to MPEG-7: Multimedia Content Description Interface*. JohnWiley and Sons, 2002.

[8] T. Sikora. The MPEG-7 visual standard for content description-an overview. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6):696 – 702, June 2001.

[9] S. Staab and R. Studer. *Handbook on Ontologies*. International Handbooks on Information Systems, Springer-Verlag, Heidelberg, 2004.

[10] R. C. Veltkamp and L. J. Latecki. Properties and performance of shape similarity measures. In *0th IFCS Conf. Data Science and Classification*, volume 1, July 2006.