**Pattern Analysis and Application manuscript No.**
(will be inserted by the editor)

# Knowledge-based Detection of Events in Video Streams from Salient Regions of Activity

**Nicolas Moënne-Loccoz[1], Eric Bruno[1], Stéphane Marchand-Maillet[1]**

Computer Vision and Multimedia Laboratory, University of Geneva

24 rue du Général Dufour 1211 Geneva 4, Switzerland

email: nicolas.moenne-loccoz@cui.unige.ch

The date of receipt and acceptance will be inserted by the editor

**Abstract** Visual events occurring in video streams (such as human postures or more complex activities) are detected from a robust and generic region-based representation of the visual content and inferred using a spatio-temporal language that integrates domain-specific knowledge. More specifically, salient regions of activity are first extracted from the dynamic of the salient points along the scene. They are mapped to a vocabulary of the domain, using a state-of-the-art classifier, to describe the visual content in terms of semantic facts. Occurrences of events, modelled as assertions of a language representing spatio-temporal relationships between facts, are in-

ferred from the description of videos by applying a forward reasoning engine. An application to visual events retrieval in videos of meetings is presented as a test case.

## 1 Introduction

Detection and recognition of events in video streams is the task to decide the occurrence of visual concepts in the spatio-temporal (2D+T) space of the captured scene. In order to achieve this task, two main approaches are usually used. The bottom-up approach first designs a specific set of features that are likely to characterise well a given event. Then, the event appearance is learnt using a large set of annotated examples (see [6, 11, 12]). This signal-based approach usually performs well ; it is robust and fast. However, it can not handle situations where events are interactively specified by end-users or where there is not enough examples (training data) available for the learning process. Alternatively, the top-down approach designs a semantic model of the event and uses an inference process in order to decide upon its occurrence (see [4, 15, 22]). This knowledge-based approach requires a well-defined description of the content and thus makes strong assumptions about the captured scene that drastically limits its use for other applications.

In this paper, we address the problem of detecting events using a generic representation of the visual content and avoiding the use of learning complex spatio-temporal patterns. Our contribution is therefore twofold, a generic and robust representation of the visual content is first proposed. It corre-

sponds to the 2D+T volumes of the visual scene that capture the main visual entities appearing in the scene during the sequence. Salient regions of activity are independent from the domain of application and from the events that will be queried. As a trade-off, such a description cannot be used directly to perform complex events detection and recognition. Thus, salient regions of activity are labelled according to a domain-specific vocabulary. This results in a semantic description of the content of videos that represents instances of visual concepts and their spatio-temporal relationships (facts). In parallel, an event language is defined to model and infer spatio-temporal patterns of facts. The language permits to express the main spatial and temporal relationship while remaining light in order to limit the side-effects of the noise coming from the generic low-level description, to ease its use for end-users, and to limit the complexity for the inference process. An event is queried as an assertion of this language. An inference engine is finally used to discover within the videos semantic description the patterns of facts satisfying this assertion.

The structure of the paper is as follows. Section 2 first reviews the extraction of the salient regions of activity. Section 3 presents the semantic description of videos, in terms of the facts that are associated to the extracted regions of activity. Section 4 defines the language used to express events as spatio-temporal patterns of facts and discusses the inference process. Finally, section 5 proposes as a test case, an application of the presented method to retrieve visual events within videos of meetings.

## 2 Salient regions of activity

In order to detect and recognise events occurring in video streams, a robust and generic description of the content is presented. The model is a set of moving ellipsoid regions with homogenous color. Such regions, called *salient regions of activity* [10], characterise objects or parts of objects that are moving within the scene.

Moving region extraction is usually performed using spatio-temporal segmentation ; but it is a time-consuming process, which often depends on the capturing environment (static camera, background model). With the aim of lowering the computational load while preserving accuracy, salient regions of activity are extracted here from moving multi-scale salient points with an adaptive *Mean-Shift*.

### *2.1 Salient points*

Salient points are points in the image space where the intensity changes in at least two directions. Several algorithms have been proposed to extract salient points, among which the wavelet-based salient points [19] and the multi-scale Harris corners [8] are considered as the most robust ones. In our work, we use the multi-scale interest points proposed by K. Mikolajczyk and C. Schmid in [8] because they have the desirable property to be scale invariant.

The multi-scale interest points are extracted from the scale space of an intensity image $I(v), v \in V = \{x, y\}$. The scale-space is obtained by

convolving $I(v)$ with a Gaussian derivative kernel for a set of scales $s \in S$ :

$$L_{v_i}(v,s) = I(v) \star G_{v_i}(s), \forall v \in V, \forall s \in S \tag{1}$$

where $G_{v_i}, i \in \{1,2\}$ is the Gaussian derivative along the dimension $v_i$ of the image space $V$. The scale normalised Harris function $H(v,s)$ is computed at each scale $s$ and each location $v$ :

$$H(v,s) = Det(\Sigma(v,s)) - \alpha Trace^2(\Sigma(v,s)) \tag{2}$$

$$\Sigma(v,s) = s^2 G(v,\sigma) \star \begin{pmatrix} L_{v_1}^2(v,s) & L_{v_1}L_{v_2}(v,s) \\ L_{v_1}L_{v_2}(v,s) & L_{v_2}^2(v,s) \end{pmatrix} \tag{3}$$

$H(v,s)$ gives a measure of the cornerness of the points $v$ at the scale $s$. It is based on the strength of the eigenvalues of the auto-covariance matrix $\Sigma$ of $L_v$. According to the scale-space theory [7], the characteristic scale of a point is a local maxima of the Laplacian defined by :

$$Lp(v,s) = s^2 \left| L_{v_1 v_1}(v,s) + L_{v_2 v_2}(v,s) \right| \tag{4}$$

Thus, multi-scale salient points are defined to be points that are local maxima of $H$ in the image space and local maxima of $L$ in the scale space.

Such points are extracted for each frame of the video stream. Figure 1 shows examples of salient points extracted in two different videos sequences. In the sequel, we note $W_t$ the set of salient points $w$ extracted at the frame $F_t$ and $s_w$ the characteristic scale of point $w$.

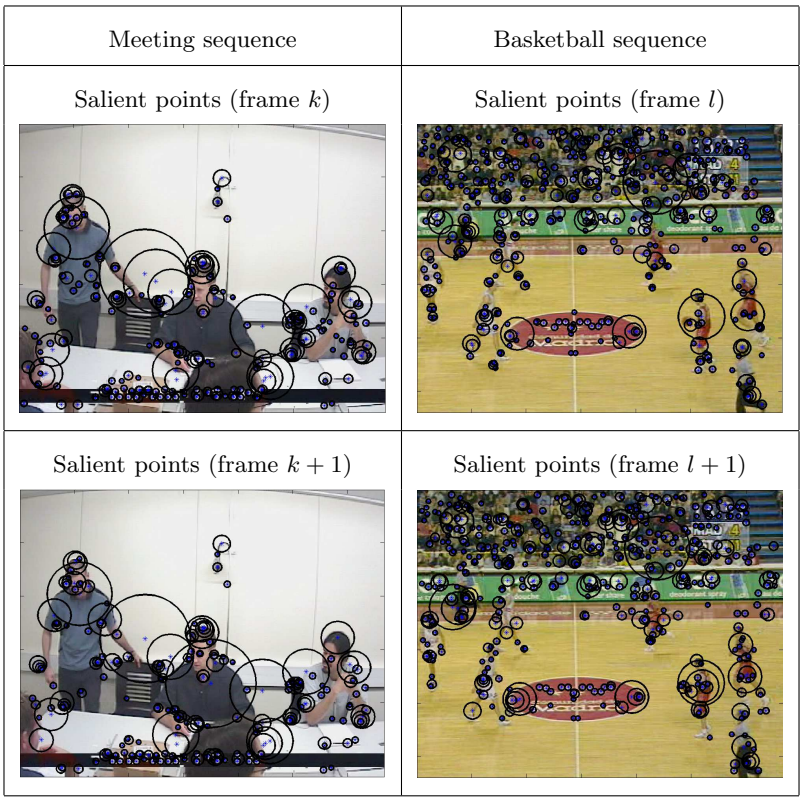


**Fig. 1** Extracted salient points for 2 successive frames (Meeting sequence from M4/IM2 corpus - Basketball sequence from MPEG-7 corpus)

### *2.2 Dynamic of salient points*

Considering a visual scene, salient points are located on the objects of interest but also on textured surfaces or non-informative background structures. Furthermore, the content of a video resides mainly within its dynamic. For these reasons, we consider only the moving salient points, i.e. points which motion is different from that of the background. In the simple case of a static camera, such points are those which position is changing over time ; but since the approach is design to make no assumption about the con-

tent, trajectories of the salient points are computed and the background motion model is estimated in order to select the relevant moving points (i.e. corresponding to moving foreground objects).

*2.2.1 Salient points trajectories* The trajectories of salient points are computed by matching pairs between two consecutive frames. More formally, for any $w_t \in W_t$, the corresponding point $w_{t-1} \in W_{t-1}$ is selected.

Whereas the *SIFT* (Scale Invariant Feature Transform) descriptors are shown in [9] to be the most robust local descriptors, we use as signature for the points in $W_t$ and $W_{t-1}$, the local gray-value invariants [17], which are differential invariants computed on the local jet of the point. The reason is that they are computationally less expensive and they provide sufficient information to match points between consecutive frames, in which case points signature remains relatively stable.

The Mahalanobis distance $d(w_t, w_{t-1})$ is computed for each pair of points. The associated covariance matrix, that integrates the noise and the correlations between dimensions of the feature space, is estimated on a large set of representative salient point signatures. In our experiment we use salient points extracted during the 5 first seconds of the video sequence ($\approx$ 20,000 points). Hence, the distance $d(w_t, w_{t-1})$ permits to estimate matches between two successive frames. Matches $M_t$ are selected, among the set of possible matches, by a greedy algorithm that tends to minimise the sum of the distances $\sum_{(w_i,w_j)\in M_t} d(w_i, w_j)$.

We therefore obtain the set of matches $M_t$ that associates the salient points of $W_t$ in the current frame with their corresponding salient point in the previous frame. The set of match $M_t$ corresponds to the set of trajectories of the points between the two successive frames $F_{t-1}$ and $F_t$. Figure 2 presents trajectories estimated for the frames whose salient points are shown in the figure 1.

*2.2.2 Global motion model estimation* Given the set of trajectories $M_t$, we estimate the most representative affine motion model (see [18] for an overview of motion estimation). This model thus corresponds to a global description of the background motion. We choose the affine motion model because of its ability to capture the main camera motions with a limited number of parameters:

$$d(v) = \begin{pmatrix} a_1 & a_2 \\ a_4 & a_5 \end{pmatrix} v + \begin{pmatrix} a_3 \\ a_6 \end{pmatrix}$$

To estimate the motion model from the set of trajectories $M_t$, we first apply a RanSaC algorithm [3], which tends to select the most representative motion model. As the set of trajectories contains noise, the model is then smoothed by applying a *Tukey* $M$-estimator in a way close to the one presented in [20]. Figure 2 shows examples of estimated motion models : for the "meeting sequence", there is no camera motion while the "basketball sequence" contains a left panning of the camera.

*2.2.3 Moving salient points* Moving salient points are those that do not follow the background motion model. In order to remove potential noise, only salient points detected as moving for a given minimal time interval are selected. Figure 2 shows some estimated moving salient points. Such points are mainly located on moving objects and provide high information about the dynamic content of the scene. However, as they are located on corner-like regions, they do not provide a suitable representation of the content. We therefore make a step ahead by estimating from them a set of *regions of activity*.

### *2.3 Regions of Activity*

Regions of activity are defined to be moving regions having an homogenous color distribution. Such regions correspond to moving objects or parts of moving objects. Moving salient points are, for the most, located on such regions. Hence, by using the characteristic scale of salient points along with the color distribution of their neighbourhood, regions of activity are extracted using a *Mean Shift* augmented with a kernel adaptation step.

*2.3.1 Adaptive Mean Shift algorithm* The Mean Shift algorithm has been used to track regions of interest [2]. The main idea is to compute an offset $\delta_{v_r}$ between a current estimation of the region location $v_r$ and an estimation $v'_r$ having a higher likelihood. The offset is computed by :

$$\delta_r = \frac{\sum_v K(v-r)p(v)(v-r)}{\sum_v K(v-r)p(v)}$$

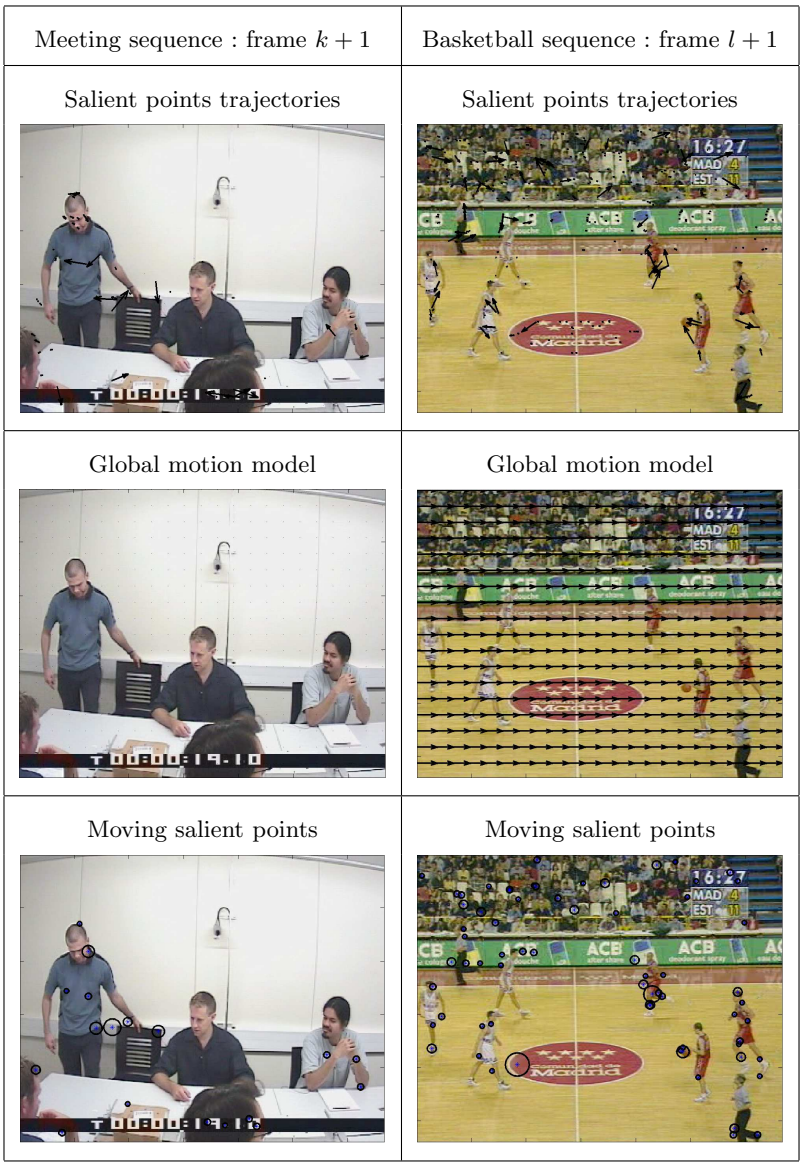


**Fig. 2** Estimated trajectories, motion models and moving salient points for the frames whose salient points are shown in figure 1

where $K$ is a kernel centred in $v_r$ and $p(v)$ is a weighting function measuring the likelihood of the pixel $v$ to belong to the region. From an initial location, the algorithm computes a new location from the offset $\delta v_r$ and is iterated until the location converges to a local maxima (mode in the representation space). In order to estimate salient regions of activity, a Mean Shift is applied from every moving salient point. As the weighting function, we use the likelihood $P(v|\theta_w) = N(\mu_{\theta_w}, \Sigma_{\theta_w})$, the probability of a pixel to be in the Gaussian $RGB$ color space defined by $\theta_w$ as follows. For a given salient point $w$, the parameters $\theta_w$ is estimated from its spatial neighbourhood at its characteristic scale (i.e., the neighbourhood inside the circle of radius $3 * s_w$ centred at the point location). The kernel is an ellipsoidal Epanechnikov Kernel :

$$K(v-r) = \begin{cases} \frac{3}{4}\left(1 - \left(\frac{(v-v_r)^T \Sigma_k (v-v_r)}{\sigma_s}\right)^2\right) & if \quad |\frac{(v-v_r)^T \Sigma_k (v-v_r)}{\sigma_s}| \leq 1 \\ 0 & otherwise \end{cases}$$

where $\Sigma_k$ is the affine transformation corresponding to the shape of the kernel and $\sigma_s$ is the size of the kernel (its radius in the affine transformed domain). The shape of the kernel is initialised to a circle, i.e. $\Sigma_k = diag(1, 1)$. Its size is set to the neighbourhood of the salient point in its characteristic scale, i.e. $\sigma_s = 3 * s_w$. However, as the shape and the size of the region is inevitably not well modelled by this arbitrary kernel, a shape adaptation step is necessary after the convergence of the Mean Shift. The shape and the size of the kernel are estimated from the covariance matrix of $P(v|\theta_w) \forall v \in r$. The algorithm then alternates a Mean Shift algorithm with an adaptation

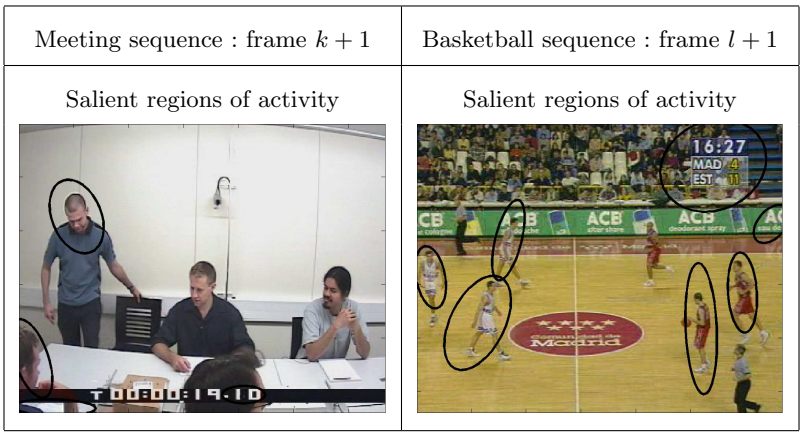


**Fig. 3** Detected salient regions of activity, estimated from the moving salient points presented in figure 2.

step until both the location of the region $v_r$ and the kernel converge. In order to speed up the process, some divergence criteria may be defined such as a maximum size of the Kernel and a minimum sum of the weights $P(v|\theta_w)$.

*2.3.2 Representation of the content* The visual content of a video stream is then represented as the set of salient regions of activity detected for each frame (see figure 1 for examples of extracted salient regions of activity). Salient regions of activity are described by a 16-dimension vector composed of the 2 components of the region location $v_r$ , the 3 components of the kernel shape $\Sigma_k$ , the size of the kernel $\sigma_s$, the 9 parameters of the Gaussian color distribution $\theta_r$ and the frame index $t$ at which the region has been detected. The regions are also defined by the set of salient points belonging to them. The trajectories of these points are used to match regions from frame to

frame. Thus, a tracking process is not required to preserve the identity of the regions over time.

## 3 Semantic description

Salient regions of activity are likely to be moving objects or noisy regions. In order to be able to infer occurrence of events, this description has to be mapped into a well-defined semantic description. Thus, the semantic attached to the regions is learnt using a classifier. Given a domain-specific vocabulary, they are labelled according to the visual concept they represent. In that way, a semantic description of the video is provided, which may be augmented with domain knowledge. This description, noted in the sequel $V_{sd}$, corresponds to the set of labelled salient regions of activity with their spatio-temporal characteristic, that we call facts.

More precisely, a fact is a semantic spatio-temporal entity. Hence, a fact is defined by its label (concept) in a given vocabulary, its spatial region within the captured scene, and a temporal interval. Salient regions of activity correspond to the observations of such facts. Thus, using a set of annotated salient regions of activity, the correspondence between regions description and the vocabulary may be learnt to produce a description of the content in terms of facts.

*3.1 Learning of facts*

A Support Vector Machine ($SVM$) classifier [21] is used to learn the mapping between the salient regions of activity represented by a vector in a feature space and a discrete vocabulary. The feature space used to classify the regions of activity is the 10-dimension space, $(\Sigma_k, \sigma_s, \theta_r)$. $\Sigma_k$ and $\sigma_s$ correspond to the ellipsoid parameters and $\theta_r$ to the Gaussian color distribution parameters. Others descriptors may be computed from the pixels of the region in the case where more information is needed to discriminate efficiently between concepts. It is important to note that the region frame index $t$ is not a discriminative feature. Furthermore, the position of the region $v_r$, while being discriminative in some cases, may lead to over-fitting, and reduce the classification accuracy.

A set of salient regions of activity has to be annotated, and the $SVM$ classifier is trained, finding the boundaries of the concepts in the feature space. As such regions correspond to the main visual entities of the domain of interest, there is no problem on the availability of training samples.

In that way, salient regions of activity are mapped to their corresponding fact. The semantic description $V_{sd}$ of a video is thus the set of facts characterised by a concept label, an ellipsoid surface of the scene, a frame index and an identifier. The identifier is such that two facts corresponding to salient regions of activity that have been matched, have the same identifier. These facts corresponds to the same instance of a concept, observed at different time points.

*3.2 Domain Knowledge facts*

In addition to the facts corresponding to the salient regions of activity, a set of domain specific and time invariant facts may be added in order to increase the expressiveness of the description. Such facts come from expert-knowledge about the domain of interest. They apply mainly on the semantic of the scene layout.

## 4 Visual Event Language

Visual events are defined as spatio-temporal patterns of visual entities. Hence, in order to model and infer events, we define a spatio-temporal language. Existing works have addressed the definition of such languages (e.g. [1, 13]) ; they are based on different spatial and temporal logics or calculus. However, the proposed methods are always dependent on the underlying representation of the content. Furthermore, while being complete with respect to the spatio-temporal relations they support, they are sensible to the noise of the underlying visual entities extraction and are too complex to be effectively used in a systematic way by end-users. In this regard, the visual event language we propose relies on the previously defined facts, that are generic semantic visual entities, and supports only a subset of the spatial and temporal relations enriched with some features of description logics ($DL$).

*4.1 Language definition*

The proposed visual event language $L$ is defined by the tuple $\{F, V, O_d, O_s, O_t\}$ where $F$ is the set of fact variables, $V$ is the vocabulary of the language, $O_d$ are the description operators, $O_s$ are the spatial operators and $O_t$ are the temporal operators.

*Description operators* model the relations between the concept associated to the facts. Description logics are usually used to perform ontology reasoning. For the sake of simplicity, we restrict the description operators $O_d$ to the class testing operator $\sqsupseteq$ and to the instance equality operator $\equiv$ defined by :

- $\forall f \in F, \forall v \in V \quad f \sqsupseteq v \Rightarrow f$ is an instance of the concept $v$
- $\forall f, f^{'} \in F, \forall v \in V \quad f \equiv_v f^{'} \Rightarrow f$ and $f^{'}$ are the same instance of $v$

In the sequel, we will write $f \equiv f^{'}$ for $f \equiv_v f^{'}$ where $v$ is the most specific concept such that $f \sqsupseteq v$.

*Spatial operators* are used to model and reason about topological or positional relations between regions. Because of the noise arising from the extraction of the regions, the number of spatial relations that may be robustly decided upon is limited. Hence, we use only 4 different spatial operators $O_s = \{\lhd, \triangle, \boxminus, \boxdot\}$ defined the following way :

- $\forall f, f^{'} \in F \quad f \lhd f^{'} \Rightarrow f$ is to the left of $f^{'}$ in the space of the scene
- $\forall f, f^{'} \in F \quad f \triangle f^{'} \Rightarrow f$ is above $f^{'}$ in the space of the scene

- $\forall f, f^{'} \in F \quad f \boxminus f^{'} \Rightarrow f$ is touching $f^{'}$

- $\forall f, f^{'} \in F \quad f \boxdot f^{'} \Rightarrow f$ is inside $f^{'}$

The positional operators $\{\lhd, \triangle\}$ define the relative position of the centroid of the regions. The topological operator $\boxminus$ specify that regions share a common surface, while the operator $\boxdot$ models the inclusion of the centroid of a region inside the surface of another one. In that way, the model remains robust and expressive enough to describe the main spatial relations.

*Temporal operators* are widely used to model temporal patterns. We use the main features of temporal logics, adapted to our discrete and noisy environment, thus defining the temporal operators $O_t = \{\leftrightharpoons, \rightarrow, \mapsto\}$ :

- $\forall f, f^{'} \in F \quad f \leftrightharpoons f^{'} \Rightarrow f$ occurs during $f^{'}$

- $\forall f, f^{'} \in F \quad f^{'} \rightarrow f \Rightarrow f$ occurs after $f^{'}$

- $\forall f, f^{'} \in F \quad f^{'} \mapsto f \Rightarrow f$ occurs just after $f^{'}$

The $\leftrightharpoons$ operator models the co-occurrence of facts. The $\rightarrow$ operator defines the relative temporal position of facts and the $\mapsto$ operator add to the position operator the constraint that no equivalent fact occurs between the two.

### *4.2 Language formulae*

A formula of the language is an association of variables $f \in F$ and terms of the vocabulary $v \in V$ with the different operators. In order to constraint the set of possible formulae, we adopt the taxonomy proposed in [16] that

classifies events into 3 categories : *states*, *events* and *scenarios*. Hence, a well-formed formula of the language $\Theta$ is either a *state* formula, an *event* formula or a *scenario* formula.

*States* formulae are identity and spatial patterns of visual entities in given intervals of time : $\forall o_s \in O_s,\ f \in F,\ v \in V$

$$\begin{aligned} \Theta_{st} &:= f \mid \Theta'_{st} \sqsupseteq v \mid \Theta'_{st} \equiv_v \Theta''_{st} \\ &:= \Theta'_{st}\ o_s\ \Theta''_{st} \mid \neg\Theta'_{st} \mid \Theta'_{st} \wedge \Theta''_{st} \end{aligned} \tag{5}$$

*Events* are the evolution of some *states* overtime. Thus an *event* formula is a temporal pattern of *state* formulae : $\forall o_t \in O_t$

$$\Theta_e := \Theta_{st}\ o_t\ \Theta'_{st} \tag{6}$$

*Scenarios* are temporal patterns of *events*. A *scenario* formula is thus defined as : $\forall o_t \in O_t$.

$$\Theta_{sc} := \Theta_e \mid \Theta_e\ o_t\ \Theta'_e \mid \neg\Theta_e \mid \Theta_e \wedge \Theta'_e \tag{7}$$

*4.3 Event inference*

As stated before, the semantic description $V_{sd}$ of the visual content of a video sequence is the set of facts occurring within the captured scene. An *assertion* of the language is defined as :

$$\Phi : \{f\} \subset V_{sd} \models \Theta \tag{8}$$

Such an assertion is satisfied when the formula $\Theta$ is valid under the context of the semantic description of a video $V_{sd}$. We will simplify the notation by equivalently writing $\Phi(\{f\}) : \Theta$.

In addition, we add the possibility to infer formula from the non-existence of facts :

$$\Phi \quad : \quad \{f\} \subset V_{sd}, \{\bar{f}\} \nsubseteq F \quad \vDash \Theta \tag{9}$$

Again, we will simplify such assertion by writing $\Phi(\{f\}) :\backsim \bar{f} \quad | \quad \Theta$.

Event assertions correspond to event queries or knowledge about the domain added by experts in order to increase the abstraction level of the model.

A forward-chaining rule-based expert system is used to infer within the semantic description of a video, the spatio-temporal patterns of facts that satisfy the queried event assertion. The CLIPS expert system (see [5] for details) is run against the semantic description $V_{sd}$ of a video, added with the domain knowledge facts. This way, using the pattern matching capabilities of CLIPS, occurrences of state, event or scenarios are retrieved.

## 5 Application

We apply the complete method in the context of visual events retrieval in videos of meetings (see [14] for an overview of the domain).

Main objects in a meeting are the meeting participants. Thus, the salient regions extracted represent, for the most, the body parts of the participants. The following vocabulary is defined : {*Head, Hand, Body, Noise*}. Table 1


**Table 1** Examples of salient regions of activity and the associated labels of the vocabulary

| | Head | Hand | Body | Noise |
|---|---|---|---|---|
| Head | 0.9091 | 0 | 0 | 0.0909 |
| Hand | 0 | 1.0000 | 0 | 0 |
| Body | 0 | 0 | 1.0000 | 0 |
| Noise | 0.0526 | 0 | 0 | 0.9473 |

**Table 2** Confusion Matrix of the learning process

presents examples of extracted salient regions of activity with their corresponding label. The mapping between regions and concepts of the vocabulary is learnt using a SVM classifier with a polynomial kernel of degree 5, and a set of 500 annotated sample regions. Table 2 presents the resulting confusion matrix. As expected the class *noise* is not well-identified because it corresponds to an ill-defined region of the feature space.

The following set of expert-knowledge facts are defined : {*Door, Board, SittingSpace*}. It corresponds to the possible locations within the meeting room (see figure 4 to view the layout of the room). In addition the following


**Fig. 4** Representation of the layout of the scene as defined by the expert knowledge.

expert-knowledge assertions are specified :

$$\begin{aligned}
\textbf{in-room}(f) &: f \sqsupseteq Head. \\
\textbf{sit}(f) &: \textit{in-room}(f) \wedge (f \boxdot \textit{SittingSpace}) \\
\textbf{standing}(f) &: \textit{in-room}(f) \wedge \neg(f \boxdot \textit{SittingSpace}) \\
\textbf{near-board}(f) &: \textit{in-room}(f) \wedge (f \boxdot \textit{Board}) \\
\textbf{near-door}(f) &: \textit{in-room}(f) \wedge (f \boxdot \textit{Door})
\end{aligned}$$

For the evaluation, we consider 5 different visual events : {*sit-down, stand-up, to-board, enter, leave*}. They may be queried by end-users using the following assertions :

$$\textbf{sit-down}(f,\ f^{'}) : (standing(f^{'}) \mapsto sit(f)) \wedge (f \equiv f^{'})$$

$$\textbf{stand-up}(f,\ f^{'}) : (sit(f^{'}) \mapsto standing(f) \wedge (f \equiv f^{'})$$

$$\textbf{to-board}(f,f^{'}) : (\neg near\text{-}board(f^{'}) \mapsto near\text{-}board(f)) \wedge (f \equiv f^{'})$$

$$\textbf{enter}(f) : \backsim \bar{f} \mid (in\text{-}room(\bar{f}) \mapsto near\text{-}door(f)) \wedge (\bar{f} \equiv f)$$

$$\textbf{leave}(f) : \backsim \bar{f} \mid (near\text{-}door(f) \mapsto in\text{-}room(\bar{f})) \wedge (\bar{f} \equiv f)$$

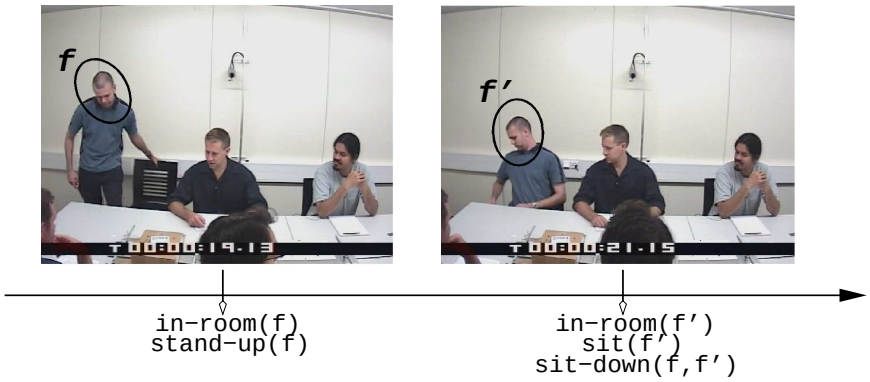


**Fig. 5** Events inference in a meeting sequence : assertions that are satisfied by the facts $(f, f^{'})$ correspond to the detected events.

A complete meeting has been annotated ; it corresponds to approximately 20 minutes of videos (35,000 frames). Occurences of events are depicted in Table 3. The patterns of facts satisfying the event assertions, that are found by the inference engine in the set of videos descriptions, are compared to the corresponding annotations. Figure 5 shows an example of the inference of an event : the semantic description of the segment has been found to satisfy the *sit-down* assertion. Results for the whole meeting are presented in Table 3.

| Events | Occurences | Precision | Recall |
|---|---|---|---|
| sit-down | 5 | 0.43 | 1 |
| stand-up | 6 | 0.5 | 1 |
| to-board | 2 | 1 | 1 |
| enter | 3 | 0.2 | 1 |
| leave | 4 | 0.25 | 0.5 |

**Table 3** Number of Occurences, Precision and Recall for each queried event

First of all, it should be noted that the number of occurences of the considered events prevents the use of a learning approach which requires a high number of training samples. Considering precision and recall values, it may be observed that false detections occur. This is due to several reasons. First, the noise affecting the extraction of regions and resulting in the misclassification of these regions is propagated to the inference process and causes some false detections. However, the main reason is that assertions may not model correctly the corresponding event. For example, the *enter* and *leave* events show low precision values. This is because meeting participants may sit inside the *Door* space (*SittingSpace* and *Door* share a common region of the scene). As the head of participants often stops moving, the head is not anymore detected and reappear only when it moves again. Hence, an *enter* event or a *leave* event is inferred with respect to the corresponding assertion. By redefining the assertions, taking into account such collisions of events, better results may be obtained. However, the recall values show

that the approach is suitable to retrieve occurences of complex events in video sequences.

## 6 Conclusion

We have presented a generic knowledge-based approach to visual event retrieval in video streams. Salient regions of activity have been defined and qualitatively shown to provide a robust and generic representation of the visual content of video streams. One specificity of our contribution is that the extraction of these regions makes no assumption on the domain of application and thus may be applied in any context. To retrieve events from the salient regions of activity, a semantic description is produced by mapping these regions onto a domain-specific vocabulary. Then, events are retrieved (modelled as assertions of a spatio-temporal language) using a forward-chaining production rule system. The validity of the method has been demonstrated through an application in the meeting domain.

Events have been successfully retrieved from combining the knowledge of the domain and the facts corresponding to labelled salient regions of activity. The proposed event retrieval approach is a good alternative to appearence learning approaches, especially in the case when events are to be defined by end-users or where there is not enough samples to train a classifier.

Future work will consider the integration of fuzziness in the visual event language to increase the robustness of the approach. Also, the language will be extended to supported richer spatio-temporal relations between facts.

This way, we expect to decrease the impact of noise while increasing the expressivness of the query language in order to achieve better retrieval performances.

## 7 Acknowledgments

This work is funded by EU-IST project M4 (www.m4project.org) and the Swiss NCCR IM2 (Interactive Multimodal Information Management).

## References

1. Alberto Del Bimbo and Enrico Vicario. Symbolic description and visual querying of image sequences using spatio-temporal logic. *IEEE transactions on knowledge and data engineering*, 7(4), 1995.
2. D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *Computer and Pattern Recognition*, volume 2, 2000.
3. Martin A. Fisher and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. In *Communications of the ACM*, 1981.
4. Mallik Ghallab. On chronicles : Representation, on-line recognition and learning. In *5th Internationnal Conference on Principles of Knowledge Representation and Resonning*, 1996.
5. J. Giarratano and G. Riley. *Expert System : Principles and Programming*. 1998.
6. A.J. Howel and Hilary Buxton. Active vision techniques for visually mediated interaction. In *Image and Vision computing*, 2002.

7. Tony Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2), 1998.
8. Krystian Mikolajczyk and Cordelia Schmid. Indexing based on scale invariant interest points. In *8th Internationnal Conference on Computer Vision*, 2001.
9. Krystian Mikolajczyk and Cordelia Schmid. A performance evaluation of local descriptors. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2003.
10. Nicolas Moënne-Loccoz, Eric Bruno, and Stéphane Marchand-Maillet. Video content representation as salient regions of activity. In *International Conference on Image and Video Retrieval*, 2004.
11. Nicolas Moënne-Loccoz, François Brémond, and Monique Thonnat. Recurrent bayesian network for the recognition of human behaviors from video. In *3th International Conference On Computer Vision Systems*, 2003.
12. Nuria Oliver and Alex Pentland. Graphical models for driver behavior recognition in a smartcar. In *Proceedings of IEEE Conference on Intelligent Vehicles*, 2000.
13. I. Ounis and T.W.C. Huibers. A logical relational approach for information retrieval indexing. In *19th Annual BCS-IRSG Colloquium on IR Research*, 1997.
14. V. Pallotta, A. Ballim, S. Marchand-Maillet, and A. Lisowska. Towards meeting information systems: Meeting knowledge management. In *6th International Conference on Enterprise Information Systems*, 2004.
15. Claudio Pinhanez and Aaron Bobick. Human action detection using PNF propagation of temporal constraints. In *M.T.T. Media Laboratory Perceptual Section Report No 423.*, 1997.

16. Nathanael Rota and Monique Thonnat. Video sequence interpretation for visual surveillance. In *3rd IEEE International Workshop on Visual Surveillance*, 2000.

17. Cordelia Schmid and Roger Mohr. Local grayvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5), 1997.

18. C. Stiller and J. Konrad. Estimating motion in image sequences: A tutorial on modeling and computation of 2D motion. *IEEE Signal Processing*, 16, 1999.

19. Q. Tian, N. Sebe, M. S. Lew, E. Loupias, and T. S. Huang. Image retrieval using wavelet-based salient points. In *Journal of Electronic Imaging, Special Issue on Storage and Retrieval of Digital Media*, 2001.

20. Philip H. S. Torr and Andrew Zisserman. Feature based methods for structure and motion estimation. In *Workshop on Vision Algorithms*, 1999.

21. V. N. Vapnik. *Statistical Learning Theory.* John_Wiley, 1998.

22. Van-Thin Vu, François Brémond, and Monique Thonnat. Temporal constraints for video interpretation. In *15th European Conference on Artificial Intelligence*, 2002.