

Human action recognition in videos based on the Transferable Belief Model

Application to athletics jumps

E. Ramasso¹, C. Panagiotakis², D. Pellerin¹, M. Rombaut¹

¹GIPSA-lab, DIS department,
46 avenue Félix Viallet, 38031 Grenoble, France
Tel: +33 4 76 57 43 50, Fax: +33 4 76 57 47 90
e-mail: e_ramasso@yahoo.fr (Corresponding author),
rombaut@lis.inpg.fr, pellerin@lis.inpg.fr

²Department of Computer Science,
University of Crete, P.O. Box 2208, Heraklion, Greece
e-mail: cpanag@csd.uoc.gr

Received: 9-30-05 / Revised: 10-29-06 / Accepted: 3-5-07

Abstract This paper focuses on human behavior recognition where the main problem is to bridge the semantic gap between the analogue observations of the real world and the symbolic world of human interpretation. For that, a fusion architecture based on the Transferable Belief Model framework is proposed and applied to action recognition of an athlete in video sequences of athletics meeting with moving camera. Relevant features are extracted from videos based on both the camera motion analysis and the tracking of particular points on athlete's silhouette. Some models of interpretation are used to link the numerical features to the symbols to be recognized which are running, jumping and falling actions. A Temporal Belief Filter is then used to improve the robustness of action recognition. The proposed approach demonstrates good performance when tested on real videos of athletics sports videos (high jumps, pole vaults, triple jumps and long jumps) acquired by moving camera and varying view angles. The proposed system is also compared to Bayesian Networks.

Key words Human action recognition, Transferable Belief Model, Temporal Belief Filter, Moving camera.

1 Introduction

Human motion analysis is an important topic of interest in Computer Vision and Video Processing communities. Research in this domain is motivated by the diversity

of applications such as automatic surveillance [1], video indexing and retrieval [2] and human-computer interaction [3]. Human actions can be extremely various, e.g. facial expression, hand gesture, human pose and people interaction. The scientific challenge is to recognize a behavior from observations coming from multimedia features such as video, audio and text [4,5]. The global problem is to link the real world which has intrinsically an analogue nature to the human interpreted world which is symbolic [6].

In the context of video indexing and monitoring applications, human motion analysis is a means to automatically analyze videos and to cope with the increasing number of videos in databases. Low level analysis is not very useful nor relevant for a end-user who prefers high level indicators [6]. In this paper, we propose an architecture to automatically recognize high level actions based on low level shape-motion and understandable features. The database is composed of video sequences of jumps (long jumps, high jumps, pole vaults and triple jumps) and the objective is to determine athlete's actions such as running, jumping and falling. The database is made of real videos acquired by a moving camera under varying view angles and can concern indoor or outdoor meetings. Videos mainly comes from broadcast TV and are compressed. Some samples of the database are pictorially described in Figs. 1 and 9. Fig. 1 illustrates original images and tracking results (three points): white level pixels correspond to the detection of human and grey level to noise (due to other moving objects and athletes).

Architectures proposed for human motion analysis generally consists in three main steps: (i) the choice of relevant numerical *features*, (ii) the definition of *models* of symbols with respect to the features and (iii) the conclusion about the reality of the symbols obtained by a *fusion* process. Relevant *features* must be chosen from the real world and correspond to numerical measures obtained by signal and image processing (there are many technics largely based on statistical approach). They must bring information about the symbols corresponding to the human. We have chosen these features by expertise, based on basic assumptions. *Models* of interpretation are used to link features and their combinations to symbols of interest. The usual approach is a learning process from a reference database [7]. The learning process cannot be a blind one because in this case, the obtained symbols are not understandable for human interpretation. If a sufficient large indexed-base is available, a supervised learning can be applied to define the symbols models. However intensive learning is often needed [8] as for instance with hidden Markov models. Human action-based video indexing cannot be strict at any time because human behavior is intrinsically continuous. Moreover, there is a great disparity between the individuals in the realization of a same activity, and sometimes the same individual performs differently as well. On the other hand, the database, the outputs of im-

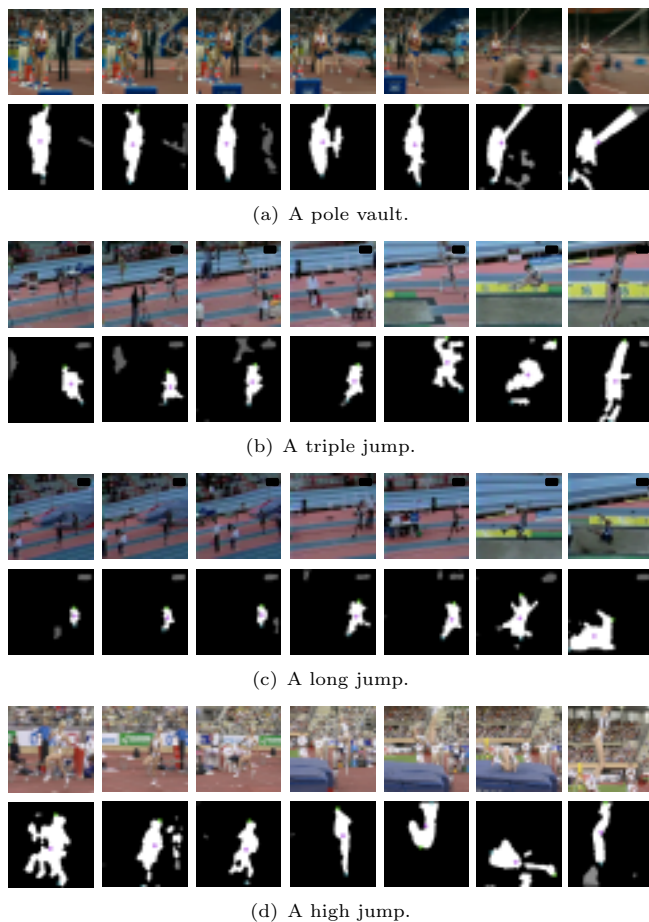


Fig. 1 Illustrations of images and tracking results on pole vault, triple jump, long jump and high jump activities. These athletics jumps videos are acquired by a moving camera with different view angle and position from the athlete. The samples show moving objects such as other athletes (e.g. in triple jump) or referee (e.g. in pole vault). White level pixels corresponds to the detection of human and grey level to noise (other moving objects).

age/signal processing and the models are not completely reliable and accurate. The symbols *recognition process*, generally made by fusion, must take into account these problems in order to make more robust the recognition. The classical approach is based on the Bayesian framework [9]. It is adapted when large databases are available but suffers from [8] intensive learning requirement, misunderstanding of the learned models and difficulty to add new information. More recent and almost unexplored (for human motion analysis) approaches [10] are the Possibility Theory (associated to the fuzzy sets and possibility measures) and the Transferable Belief Model (TBM) (based on belief functions and plausibility measures). Differences between both fusion approaches are discussed in [11] on a real example¹. Possibility is well

¹ Many other papers are available on Smets' homepage <http://iridia.ulb.ac.be/~psmets>. The web page also pro-

posed links towards other researchers in the TBM community as well as softwares.

adapted for poor information and qualitative description such as ordinal information whereas TBM is more adapted for numerical information, reinforcements and compensatory effects in combinations [11]. TBM manages smartly rules [12] produced by experts or systems and which are useful for databases management or indexing [13]. In the context of human motion analysis as concerned in this paper, we explore the application of the TBM.

The TBM has been developed by Smets [14] from the previous work of Shafer on *Evidence Theory* [15] (see [16] to analyze the differences). This theory makes it possible to take into account the continuous and blurred aspects of the human behavior such as in transitions between actions. Indeed, it allows the explicit modelling of *doubt*. Doubt, intrinsically present in human judgement and algorithm results, is useful to represent total ignorance state like missing a priori whereas probability generally assumes the equiprobability principle. Moreover, *conflict* is quantified within the TBM and this relevant information can be used for the questioning of models or rules defined beforehand by an expert but not corresponding anymore to the reality of the data. Conflict is at the core of the *Temporal Belief Filter* (TBF) developed in [17] to improve human action recognition by smoothing belief functions and separating actions states. The conflict was also used for belief functions clustering [18]. Doubt and conflict information are seldom incorporated for human motion analysis. In this paper, a new architecture for human action recognition in athletic sports videos based on the TBM is proposed. Temporal aspects of human motions are taken into account using the TBF which is one step towards (unexplored) activity recognition in the TBM framework.

The remainder of the paper is as follows. Related work is discussed Section 2, an overview of the proposed recognition architecture is presented Section 3, features extraction is described Section 4, action models and TBM framework are dealt with in Section 5, the action recognition process is detailed Section 6 and Temporal Belief Filter is presented Section 7. Experimental results are described Section 8. Finally, Section 9 is dedicated to conclusion and future work.

2 Related work

Many methods have been proposed for action recognition [19]. Generally, a recognition system consists in comparing observations to models which are generally learned from databases or set by expert knowledge. Models can be reduced to clusters and the recognition becomes a problem of classification [5]. Models can also be features vectors as in template matching [20] and the

recognition process consists in choosing the best template according to a measure of similarity. But generally, models are built from huge databases [7] under a bayesian framework [9]. Most of methods belong to this active category deserving attention.

In the probability modelling context, an action or an activity is often described by means of state machines which is an intuitive approach close to human-reasoning. State-based methods consists in building state machines and then let it evolve according to observations. Hidden Markov Models (HMM) and Dynamic Bayesian Networks (DBN) are well-known in the Computer Vision community. They have been widely studied [21,22]. Some adaptations of these methods have been also proposed notably in [23] where authors exploit DBNs, Partially Coupled HMM, whose topology is determined using the Bayesian Information Criterion, and Multi-Observation HMM for causality discovery and events modelling. In [24], an interesting description and comparison between DBN and HMM is proposed for sports video sequence interpretation. In [13], HMM, DBN and rules are integrated in one system for multimedia database management.

Neural Networks represent models based on states but are not based on probability. The network topology is determined by means of optimization procedures requiring also large learning sets. In human action and activity recognition, Time Delay Neural Networks are applied [25].

Despite huge databases, even including multimodal information [4,5], *domain-specific information* is often required to assign an application dependant semantic to the result of the recognition process [26,27]. Sometimes, systems are built only for one application [28].

The community of Artificial Intelligence has also focused on state-based models representation. One important tool is the Petri net [29]. A Petri net is able to take into account the synchronization problem. It is often used for monitoring and control as in [30] with application to nuclear power plant supervision. Many adaptations of this framework has been proposed notably for including fuzzy measures [31] and stochastic aspect [32].

Based on belief theory, a few work can be cited: in [33] a method based on rules is proposed for manoeuvre recognition and in [34], Petri nets are extended to belief theory. In [35] a classifier of human postures is presented and in [36] another classifier for emotions recognition. The two last methods are based on belief theory, not on TBM, actually they do not consider conflict and use Dempster rule. Moreover their methods are static whereas in this paper we use the Temporal Belief Filter [17] which takes temporal aspects of belief into account. The Transferable Belief Model, proposed by Smets and Kennes [14] is therefore originally exploited in this paper for human action recognition based on belief functions. The TBM allows to manage uncertainty, imprecision, expert rules, partial knowledge, conflict and open world assumption.

3 Recognition architecture

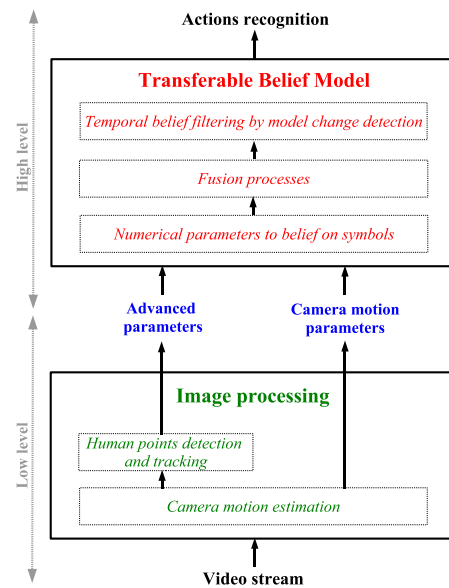


Fig. 2 The proposed architecture for human action recognition in videos. It is based on two levels of processing and relies on the TBM framework for belief representation and combination.

Human action recognition requires several steps as depicted in Fig. 2. The video stream is provided by real video sequences of athletics jumps preliminarily acquired by a moving camera. In the low level processing step, relevant features are extracted. They are generally application dependant [4] and based on more or less assumptions [37]. In this paper, the choice of the features is based on three a priori assumptions:

- The human is tracked by the cameraman. This assumption is satisfied when the human is the center of interest (e.g. in athletics jumps, athletes are tracked to satisfy telespectators, trainers and sponsors).
- A single human is moving. The case of multiple humans is more complex and not considered here.
- The trajectories of particular human body points (human's head, center of gravity and one end of leg) give information on actions. The system is generic enough to add new information.

The two first assumptions are very common in Computer Vision as discussed in the survey of Moeslund [37] (from year 1980 to 2000). In sports video, the athlete is generally the object of interest thus the first assumption holds. The system should be improved to be applied in surveillance applications where multiple people are to be considered. Usually [1], multiple people are tracked in controlled and indoor environments using color-based features [38] which are not robust to be used in real videos acquired in outdoor or indoor scene as it is the case in sports videos.

In the high level processing step, features are assigned a semantic according to actions trueness. Semantic assignment consists in describing actions by means of weighted symbols where weights (belief) are computed according to the values of features. A variety of opinions is thus available concerning actions and a consensus is obtained by combining them in the Transferable Belief Model framework. Then, a Temporal Belief Filter [17] is applied to ensure temporal consistency of the opinions as well as action transitions discovering. The proposed architecture is built such as to be generic enough to add new features and new actions.

4 Features extraction

In this section, the low level part of the architecture is described. Numerical features are extracted at each frame of the video and are provided by a camera motion estimation and a tracking algorithms. Features are: the horizontal translation, the vertical translation, the divergence, the variation of center of gravity, the alternation of legs, and the angle between horizon and human axis.

4.1 Camera motion estimation

An affine model is used to describe the camera motion. Such a model is generally sufficient for most of real video sequences. The flow vector $\vec{w}(p_i) = [u(p_i), v(p_i)]^T$ of a point p_i located at (x, y) in an image can be described by a parametrized affine motion such as:

$$\vec{w}(p_i) = \begin{bmatrix} u(p_i) \\ v(p_i) \end{bmatrix} = \begin{bmatrix} P_{hm} \\ P_{vm} \end{bmatrix} + \begin{bmatrix} P_{div} & 0 \\ 0 & P_{div} \end{bmatrix} \begin{bmatrix} x_i \\ y_i \end{bmatrix} \quad (1)$$

Only features $\{P_{hm}, P_{vm}, P_{div}\}$ are retained. They allow to consider 2D translation motion (horizontal for P_{hm} and vertical for P_{vm}) with divergence (P_{div}). The divergence is mainly used as a complementary feature of the horizontal translation for frontal motion. The computation of these coefficients are achieved by a robust iterative and multiresolution method described in [39]. The method has been implemented (motion2D software²) by the Vista Team of IRISA. The motion model proposed in [39] takes into account the global variation of illumination between two successive frames thus, it is robust to illumination changes as required for real videos. The method runs on gray level images thus is independant from color. This method was already successfully applied for dynamic content analysis [40] and video indexing [41].

A dominant motion image is obtained from the camera motion estimation. The intensity of a pixel in this image depends on its membership to the dominant motion that is assumed to be the motion of the background. Fig. 4(b) depicts such images corresponding to running,

² The software can be downloaded on <http://www.irisa.fr/Vista/Motion2D>.

jumping and falling actions for a high jump sequence. The silhouette of the athlete is in black because it does not belong to the dominant motion (foreground).

Fig. 3 depicts some examples for a high jump composed of four actions. References are given to emphasize the feature relevance according to actions.

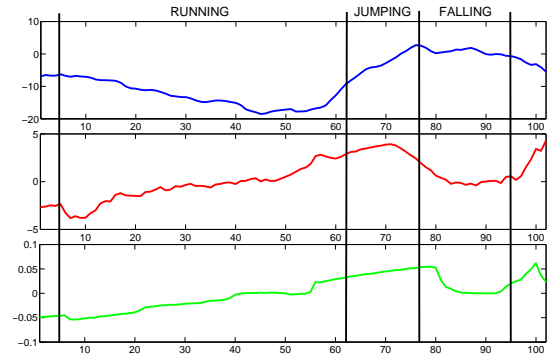


Fig. 3 Camera motion features for a high jump. From top to bottom (w.r.t frame number): P_{hm} (horizontal motion), P_{vm} (vertical motion) and P_{div} (divergence). References are given for running, jumping, falling and standing up.

4.2 Human points detection and tracking

The temporal curves of the position of the three following major human points: *head*, *center of mass* and *end of leg*, are supposed to be sufficient to help in the recognition of global actions. Among the available method [38], the human point detection and tracking algorithm presented in [42] is used here and adapted to detect and track 3 points³. The method consists of two steps: detection and tracking, and requires a binary silhouette.

4.2.1 Segmentation The dominant image motion obtained from the camera motion estimation is thresholded ($\sigma = 0.1$) and a median filter is applied to remove small regions and to create homogeneous areas. Then erosion and dilatation are combined to refine the silhouette shape.

4.2.2 Detection This initialization is executed in the first frame of the sequence. The human points detection method consists, first, of determining the center of the mass, of coordinates (x_c, y_c) , of foreground pixels. The major human body axis passing through the mass center point is then computed by calculating its orientation Θ . The orientation Θ is defined by the three

³ In the first version, 18 points are tracked but in this paper, image quality is not sufficient for such level of detail (which is not useful here).

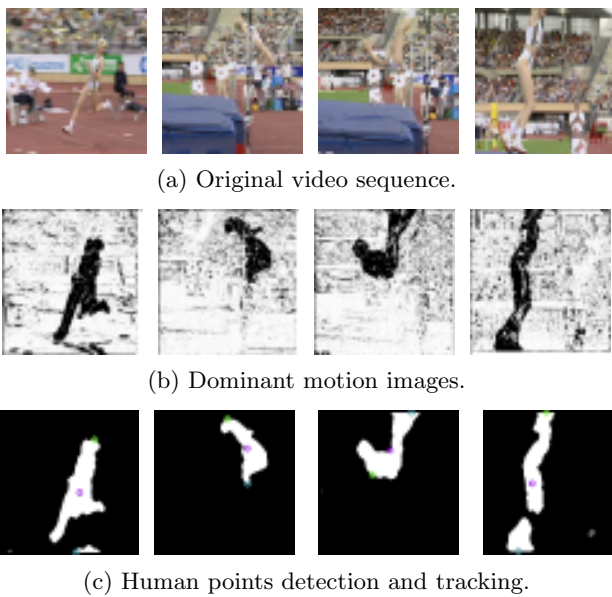


Fig. 4 Original video sequence, dominant motion images and human points detection and tracking results for a high jump.

second order central moments $C_{1,1}, C_{2,0}, C_{0,2}$ defined as $C_{p,q} = \sum_{(x,y) \in F} (x - x_c)^p (y - y_c)^q$ thus:

$$\Theta = \arctan \left(\frac{2C_{1,1}}{C_{2,0} - C_{0,2}} \right) \quad (2)$$

It is assumed that the human stands in the first frame so that the head point and the end of the leg, of coordinates (x_h, y_h) and (x_l, y_l) respectively, should be found. Given the mass center (x_c, y_c) as the reference, both previous points represent extremities of the silhouette.

4.2.3 Tracking In this step, the three points are tracked. This procedure is executed in every frame of the sequence by considering the current frame and its previous. First, the pixels of the binary silhouette image are reclassified to reduce the number of wrongly classified pixels. For that, the minimum distance of each foreground pixel from the previous position of the three human points is computed. If it is higher than a threshold (adaptive to image data and defined as a percentage of the human height) then the foreground pixel is classified as the background. Background pixels that belong to human silhouette holes are classified as the foreground class. The procedure of reclassification increases the accuracy of human points detection. In Fig. 4(c), some results are presented: white pixels correspond to moving objects (foreground), and black ones to background. Thus, the quality of the estimated human silhouette is improved. Finally, the three points are detected by the detection algorithm previously described. This method produces two pairs of solutions for the head point and the leg point, as it is unknown if the head point is found above

or under the mass center. We choose the pair which is closer to the estimated pair of the previous frame.

4.3 Synthesizing advanced features

The coordinates obtained from the tracking module are relevant but not interpretable in terms of human actions. New advanced features are computed from the coordinates in order to elucidate the description of actions:

- *Swing* (P_{swing}) describes how is positioned the main axis of the human compared to the horizontal axis (Fig. 5). This measure is an angle value and is computed in two steps: First it is assumed that the three points draw a straight and a regression based on these points allows to compute the coefficients of this straight. In the second step, the angle between the straight and the horizontal axis is computed (Fig. 5).
- *Coordinates variation* (P_{vcg}) is more relevant than coordinates itself. A high variation is interpreted as a great motion either upward or downward with respect to the sign of the variation.
- *Alternation* (P_{alter}) represents the human legs motion. The computation consists, first, in considering the axis passing through the head point and the center of gravity and, second, in detecting where is located the end of leg point (right or left hand side of the axis). The signal provided is binary (right/left) and the analysis of the frequency (using the mean of alternation in an interval of frames) allows to detect the speed of alternation.

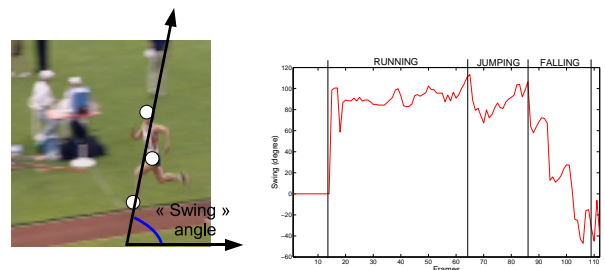


Fig. 5 Swing angle feature computation and example of variation for a high jump.

Fig. 6 depicts the three advanced features for a high jump composed of four actions. References are given to emphasize the relevance of the features according to actions.

5 Models of action interpretation

Each low level redundant/complementary features, provided here by the camera motion estimation and the

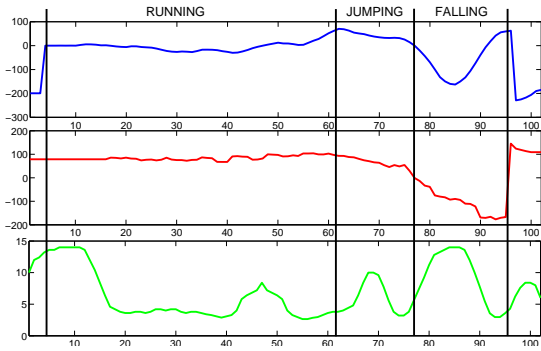


Fig. 6 Example of advanced features for a high jump sequence. From top to bottom: P_{vcg} (variation of the centre of gravity), P_{swing} (swing, in degree) P_{alter} (gait period). References are given for running, jumping, falling and standing up. The video corresponds to the same as in Figs. 3-4.

Table 1 Raw features provided by low and mid level modules.

camera motion (affine motion features)	
P_{hm}	horizontal translation
P_{vm}	vertical translation
P_{div}	divergence
tracking (coordinates)	
P_{vcg}	variation of center of gravity
P_{alter}	alternation
P_{swing}	angle between horizon and human axis

tracking, informs about the trueness of the different possible high level actions made by the human, e.g. running, jumping or falling. A semantic is assigned to each feature by means of symbols. According to their numerical value, a belief in these symbols is computed using fuzzy-inspired models of interpretation. Reliability factors are automatically computed. Then, beliefs are combined using the Transferable Belief Model (TBM) framework [14] to obtain a more complete information about actions by taking *imprecision*, *uncertainty*, *reliability* and *conflict* concerning features into account. The TBM is an axiomatically well-founded framework which relies on Evidence Theory [10] and based on the work of Shafer [15] and allows to combine distinct⁴ sources of belief. The TBM was successfully used for many applications such as detection of submarines [44] and target identification [12].

The TBM framework is well-adapted for action recognition notably because doubtful transitions between actions are explicitly modelled and conflict between features is emphasized reflecting the need to improve the fusion process.

⁴ The notion of distinctness is close but not equivalent to the independence notion in probability theory. See [43] for more details.

5.1 Numerical features to belief on symbols

5.1.1 Frame of discernment In the TBM framework, the value of a low level feature has to be converted into belief on symbols that describe an action state. This numeric to symbolic conversion shortens the semantic gap. The name of the feature is used as subscript of P . For instance, concerning the horizontal motion hm , the symbols associated to the numerical feature P_{hm} would be *small*, denoted S_{hm} , and *high*, denoted H_{hm} . In the sequel, the word *symbol* is replaced by *hypothesis* to agree with formalism in belief theory.

All hypotheses concerning a feature P are gathered in a frame of discernment Ω_P . A frame of discernment is referred as FoD in the sequel. For instance, $\Omega_{hm} = \{S_{hm}, H_{hm}\}$ is the FoD of feature P_{hm} . Each FoD is exhaustive and this supposes the *closed-world* assumption, i.e. all possible states of P are foreseen. Singletons and subsets of a FoD Ω_P are called propositions and are contained in its power set 2^{Ω_P} . For instance: $2^{\Omega_{hm}} = \{\emptyset, \{S_{hm}\}, \{H_{hm}\}, \{S_{hm}, H_{hm}\}\}$ is the set of propositions concerning the state of feature P_{hm} . A proposition composed of several hypotheses explicitly models the doubt between these hypotheses. In the sequel, and to simplify the notation, a subset of one element is replaced by this element while a subset of two or more elements is replaced by the logical OR (\cup) of these elements. For instance, $\{S_{hm}\} \leftrightarrow S_{hm}$ and $\{S_{hm}, H_{hm}\} \leftrightarrow S_{hm} \cup H_{hm}$.

5.1.2 Belief masses assignment It is necessary to quantify the confidence in each proposition because numerical features are imprecise as well as the definition of their associated hypothesis. The basic belief assignment (BBA), $m_P^{\Omega_P}$, is a belief function that assigns such weights. The mass $m_P^{\Omega_P}(X)$ is the belief on proposition $X \subseteq \Omega_P$ w.r.t. the value of feature P . The superscript is very important in order to avoid combining masses defined on different frames as we will see further. The subscript allows to distinguish the features. A BBA is defined formally as:

$$m_P^{\Omega_P} : \begin{array}{l} 2^{\Omega_P} \rightarrow [0, 1] \\ X \rightarrow m_P^{\Omega_P}(X) \end{array} \quad \begin{array}{l} m_P^{\Omega_P}(\emptyset) = 0 \\ \sum_{X \subseteq \Omega_P} m_P^{\Omega_P}(X) = 1 \end{array} \quad (3)$$

A value $m_P^{\Omega_P}(X)$ expresses a confidence in proposition $X \subseteq \Omega_A$ but does not imply any additional claims regarding subsets of X [10]. It is the fundamental difference with probability theory.

In this paper, a fuzzy set-inspired method is used to define the BBAs. Fuzzy intervals are well adapted to represent the description of a state which is inherently vague as it is the case for a feature state or an action. A trapezoidal fuzzy interval describing the proposition $X \subseteq \Omega$, with $|\Omega| = 2$, is defined by a set of four thresholds $\{th_X^1, th_X^2, th_X^3, th_X^4\}$, with $[th_X^1, th_X^4]$ representing the support of X ($m^{\Omega}(X) \neq 0$) and $[th_X^2, th_X^3]$ the core ($m_P^{\Omega_P}(X) = 1$).

Table 2 Application of the coefficients of reliability.

	P_{hm}	P_{vm}	P_{div}	P_{vcg}	P_{swing}	P_{alter}
α_{dist}				✓	✓	✓
α_{sup}	✓	✓	✓			

The thresholds setting is currently done in two steps. We assume some videos are annotated at each frame by one of the propositions concerning feature state (*high, low...*) Given these references, the mean of each feature is computed over the videos. These mean values allows to estimate the position of the trapezes. Then, in a second refinement step, the thresholds are adjusted in case they involve conflict during the fusion process. The modification consists in increasing the core of doubt. This allows to have a coherent fusion process. An automatic learning could be performed using for instance an EM approach [45].

An example of BBA concerning the horizontal motion P_{hm} is given in Fig. 7. The core of the trapezoidal fuzzy interval representing the set $S_{hm} \cup H_{hm}$ is interval $[th_X^2, th_X^3] = [4, 5]$ and the support is interval $[th_X^1, th_X^4] = [2, 6]$. For instance, if $P_{hm} = 2.3$ then BBA is $m_{P_{hm}}^{\Omega_{hm}}(H_{hm} \cup S_{hm}) = 0.33$ and $m_{P_{hm}}^{\Omega_{hm}}(S_{hm}) = 0.67$ (null for other propositions).

Fig. 7 A basic belief assignment based on fuzzy rules for the feature P_{hm} estimated by the camera motion estimator. The absolute value is taken because only the amplitude is interesting here. A belief is assigned to each hypothesis and set of hypotheses of the FoD $\Omega_{P_{hm}}$.

This method is applied for all features. It can be noticed that doubt between hypotheses explicitly appears as for instance $S_{hm} \cup H_{hm}$.

5.2 Integrating reliability of features

In the TBM, the discounting process [15] weighs the belief of a feature according to the reliability of the corresponding source. The reliability is an important tool for action recognition in video because it allows to give a penalty on belief provided by sources that work in non-optimal conditions.

A coefficient of reliability, denoted $\alpha \in [0, 1]$, is applied on a belief $m_P^{\Omega_P}$ and a new belief m_P^{α, Ω_P} is obtained as follows:

$$\begin{aligned} m_P^{\alpha, \Omega_P}(X) &= (1 - \alpha) \cdot m_P^{\Omega_P}(X), \forall X \subsetneq \Omega_P \\ m_P^{\alpha, \Omega_P}(\Omega_P) &= \alpha + (1 - \alpha) \cdot m_P^{\Omega_P}(\Omega_P) \end{aligned} \quad (4)$$

and $(1 - \alpha)$ is the dual of the reliability called discounting factor. Expert knowledge or statistics can be used to

compute this coefficient [46]. Our methodology consists in computing them from data at each frame⁵. It allows to take into account reliability that evolves w.r.t. the quality of the video. Two coefficients have been computed, one for tracking (α_{dist}) and one for camera motion estimation (α_{sup}):

- α_{dist} : the distance between the center of gravity and the head is assumed to be constant between two successive frames. The distance is normalized into $[0, 1]$ (by using the size of the image) and is used as a coefficient of reliability. When the distance is constant, the coefficient is close to 1 so the reliability is high and vice-versa. This coefficient reflects the quality of the tracking: when other moving objects appear, the binary silhouette can be of bad quality and so does the tracking.
- α_{sup} : the camera motion estimation allows to generate the support size which is a number between $[0, 1]$ reflecting the number of pixels belonging to the dominant motion. When this number is close to 1 then almost all pixels belong to the dominant motion whereas none object is moving when it is close to 0. This feature is defined by a fuzzy interval with core $[0.7, 0.8]$ and $[0.6, 0.90]$ as support. This coefficient allows to discount the features coming from the camera motion estimation.

Coefficients are sum up in Tab. 2 which indicates as well the discounted features.

6 Belief fusion for action recognition

BBAs associated to the features are now available. The objective of the fusion process is to combine these BBAs to obtain a confidence about the trueness of actions. We recall that a belief mass $m_P^{\Omega_P}(X)$ is the belief on proposition $X \subseteq \Omega_P$ according to the value of feature P . The superscript is very important in order to avoid combining masses defined on different frame as we will see in this section. The subscript allows us to distinguish the features.

6.1 Action Representation

The features do not give directly information on an action. For instance, an action can be described as follows:

IF the variation of the centre of gravity position is high AND the horizontal motion is small THEN action jumping is right

More generally, each action is described by a rule with the following prototype: IF [*condition*] THEN [*conclusion*], where a *condition* is a *logical rule* involving features

⁵ Here, the frame is an image but not a FoD.

states and the *conclusion* is the consequence on the action states. Logical rules are well handled in the TBM framework using rules of combination⁶. However, before combining the BBAs associated to the features, they must be defined on a common FoD. The fusion process is decomposed in three steps: (i) a refinement process (where BBAs are defined on product spaces), (ii) a fusion process (where BBAs are combined) and (iii) a coarsening process (where the trueness of action is inferred).

6.2 First step: refinement process

For two features, P_1 and P_2 , the cartesian product of their FoD, $\Omega_{P_1, P_2} = \Omega_{P_1} \times \Omega_{P_2}$, allows to obtain a common FoD. One says that Ω_{P_1} and Ω_{P_2} are extended to Ω_{P_1, P_2} . The extension is denoted \uparrow , e.g. $\Omega_{P_1} \uparrow \Omega_{P_1, P_2}$. The BBAs $m_{P_1}^{\Omega_{P_1}}$ and $m_{P_2}^{\Omega_{P_2}}$ must be rewritten on this common FoD before combination. For that purpose, the vacuous extension [14, 12] is applied leading to the BBAs $m_{P_1}^{\Omega_{P_1} \uparrow \Omega_{P_1, P_2}}$ and $m_{P_2}^{\Omega_{P_2} \uparrow \Omega_{P_1, P_2}}$. The vacuous extension of $m_{P_i}^{\Omega_{P_i}}$ on Ω_{P_i, P_j} (notation for $\Omega_{P_i} \times \Omega_{P_j}$) is:

$$m_{P_i}^{\Omega_{P_i} \uparrow \Omega_{P_i, P_j}}(C) = \begin{cases} m_{P_i}^{\Omega_{P_i}}(B) & \text{if } C = B \times \Omega_{P_j} \\ & \text{and } B \subseteq \Omega_{P_i} \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

For instance, feature P_{vcg} , concerning the center of gravity position, is described by the states *high* (H_{vcg}), *middle* (M_{vcg}) and *low* (L_{vcg}). The associated FoD is $\Omega_{vcg} = \{H_{vcg}, M_{vcg}, L_{vcg}\}$. The feature P_{hm} concerning the horizontal motion is described by the states *small* (S_{hm}), *high* (H_{hm}) leading to the FoD $\Omega_{hm} = \{S_{hm}, H_{hm}\}$. The cardinality of $\Omega_{hm, vcg}$ is the product of the cardinality of each FoD: $|\Omega_{hm, vcg}| = |\Omega_{hm}| \times |\Omega_{vcg}|$ equals to 6 in this example with: $\Omega_{vcg, hm} = \{H_{vcg} \cap S_{hm}, M_{vcg} \cap S_{hm}, L_{vcg} \cap S_{hm}, H_{vcg} \cap H_{hm}, M_{vcg} \cap H_{hm}, L_{vcg} \cap H_{hm}\}$ where \cap corresponds to the logical AND. The resulting BBA $m_{hm}^{\Omega_{hm} \uparrow \Omega_{hm, vcg}}$ is rewritten by applying the vacuous extension:

$$\begin{aligned} m_{hm}^{\Omega_{hm} \uparrow \Omega_{vcg, hm}}(\Omega_{vcg} \cap S_{hm}) &\Leftarrow m_{hm}^{\Omega_{hm}}(S_{hm}) \\ m_{hm}^{\Omega_{hm} \uparrow \Omega_{vcg, hm}}(\Omega_{vcg} \cap H_{hm}) &\Leftarrow m_{hm}^{\Omega_{hm}}(H_{hm}) \\ m_{hm}^{\Omega_{hm} \uparrow \Omega_{vcg, hm}}(\Omega_{vcg} \cap \Omega_{hm}) &\Leftarrow m_{hm}^{\Omega_{hm}}(\Omega_{hm}) \end{aligned} \quad (6)$$

where $\Omega_{vcg} \cap X = (H_{vcg} \cup M_{vcg} \cup L_{vcg}) \cap X = (H_{vcg} \cap X) \cup (M_{vcg} \cap X) \cup (L_{vcg} \cap X)$. That means the feature P_{hm} gives no information about the symbolic state of vcg , giving a sense to the term *vacuous* to describe the extension to the product space. This technic must be performed for all the features used in the rule's premise.

6.3 Second step: fusion process

The conjunctive rule of combination [14] (called \odot -rule), which is commutative and associative, and the disjunctive rule of combination [47] (called \oplus -rule) are generally used to combine distinct pieces of evidence.

Given two distinct BBAs $m_{P_1}^{\Omega}$ and $m_{P_2}^{\Omega}$ defined on the same FoD Ω (for instance $\Omega = \Omega_{P_1} \times \Omega_{P_2}$), their combination is defined as:

$$m_{P_1}^{\Omega} \odot m_{P_2}^{\Omega}(E) = \sum_{C \Delta D = E} m_{P_1}^{\Omega}(C) \cdot m_{P_2}^{\Omega}(D) \quad (7)$$

with $\Delta = \cap$ (resp. \cup) for the conjunctive (resp. disjunctive) rule of combination. The rules of combination are used in logical rules. Tables of rules can also be used [48].

The cardinality of a FoD Ω of the resulting BBA can be of great cardinality which can be reduced using the method proposed in [49]. Hereafter, the FoD Ω is reduced using a *coarsening process* (a kind of projection): each element in 2^{Ω} is interpreted as one elements of 2^{Ω_A} . This allows to obtain the trueness on the action A .

6.4 Third step: coarsening process

After combining the new BBAs, a coarsening operator, denoted \downarrow , based on the logical rules is applied from the FoD Ω_{P_1, P_2} to the FoD Ω_A in order to give information about the trueness of action A . The coarsening can be viewed as a mapping denoted ρ and defined as follows:

$$\rho : 2^{\Omega_{P_1}} \times 2^{\Omega_{P_2}} \rightarrow 2^{\Omega_A} \\ (X, Y) \mapsto Z \quad (8)$$

For the example given in the beginning of the part 6.1, the function ρ is described by the table of rules given Tab. 3. Action jumping A is true if H_{vcg} and S_{hm} are true, and false for the other cases. If there is any doubt about H_{vcg} and S_{hm} , it is reported to the trueness of A . It can be noticed that Tab. 3 is built with only the focal elements⁷ of the BBAs. It allows to decrease the number of element from $2^6 = 64$ to 24 elements. Tab. 3 shows 15 elements because the belief on the empty set is null by construction (eq. 3) besides, $m(L_{vcg} \cup H_{vcg}) = 0$ due to the modelling by fuzzy intervals. Furthermore, and in order to simplify the notation, the vacuous extension is not written but it is implicit.

Conclusively, the BBA $m_{P_1, P_2}^{\Omega_A}$, concerning an action A and taking into account the features initially defined on different FoDs, is defined as follows:

$$m_{P_1, P_2}^{\Omega_A}(Z) = \sum_{Z = \rho(X, Y)} m_{P_1, P_2}^{\Omega_{P_1, P_2}}(X \cap Y) \quad (9)$$

⁷ A focal element corresponds to a proposition for which the belief is not null.

⁶ De Morgan's algebra can also be applied.

Table 3 Example of coarsening by means of a table of rules. For instance, action jumping A is true if H_{vcg} and S_{hm} are true, and false for the other cases. This table is built only with the propositions for which the belief is not null. Furthermore, to simplify the notation, the vacuous extension is not written but it is implicit.

	H_{vcg}	$(H_{vcg} \cup M_{vcg})$	M_{vcg}	$(M_{vcg} \cup L_{vcg})$	L_{vcg}
H_{hm}	F_A	F_A	F_A	F_A	F_A
$H_{hm} \cup S_{hm}$	$R_A \cup F_A$	F_A	F_A	F_A	F_A
S_{hm}	R_A	$R_A \cup F_A$	F_A	F_A	F_A

According to the previous example, the following BBA is obtained:

$$\begin{aligned}
 m_{P_1, P_2}^{\Omega_A}(R_A) &= m_{P_1, P_2}^{\Omega_{P_1, P_2}}(S_{hm} \cap H_{vcg}) \\
 m_{P_1, P_2}^{\Omega_A}(R_A \cup F_A) &= m_{P_1, P_2}^{\Omega_{P_1, P_2}}((S_{hm} \cup H_{hm}) \cap H_{vcg}) \\
 &\quad + m_{P_1, P_2}^{\Omega_{P_1, P_2}}(S_{hm} \cap (H_{vcg} \cup M_{vcg}))
 \end{aligned}$$

and $m_{P_1, P_2}^{\Omega_A}(F_A)$ is the sum of all the other elements.

6.5 Note on computational issues

The use of the matrix notation in the TBM operations elucidates the transfers of belief between sets involved in rules of combination [50]. However it is not well suited for a FoD Ω of high cardinality since the matrix dimension is of $2^{|\Omega|} \times 2^{|\Omega|}$. In [51], authors proposed bit wise representation which is more effective. The difficulty of combination computation increases exponentially with the cardinality of the common FoD and some methods exist to reduce it [49].

7 Temporal belief filter

The Temporal Belief Filter (TBF) was proposed in a previous work described in [17]. The TBF works on each action independently taking as input the BBA obtained after features fusion. The TBF provides a BBA without conflict, temporally consistent (without high variation) and consonant (action states are made exclusive). The latter properties is an important characteristic of the TBF because the BBA has only two focal sets⁸: either R_A and $R_A \cup F_A$, or F_A and $R_A \cup F_A$. In the former case, the action is said to be in the *right state* while *false state* in the latter case.

The general principle of the TBF is depicted in Fig. 8. The core of the TBF is based on *implication rules* well-managed in the TBM framework [12]. An implication rule is generally used to specialize a BBA. We have interpreted implication rules \mathcal{R} and \mathcal{F} as models of evolution denoted $\mathcal{M} \in \{\mathcal{R}, \mathcal{F}\}$. Each one focuses on one hypothesis of the FoD of an action A which is either R_A or F_A . At each frame f , the TBF works in two steps: (i) state prediction and (ii) state updating.

⁸ A focal set in a BBA is a set for which the associated belief mass is not null.

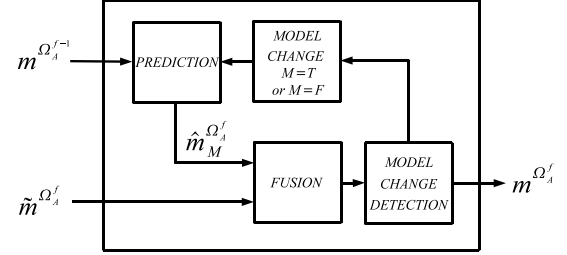


Fig. 8 The Temporal Belief Filter (TBF) principle where $\hat{m}_f^{\Omega_A}$ is the prediction, $m_f^{\Omega_A}$ is the output value of the TBF at frame f given the state of actions and $\tilde{m}_f^{\Omega_A}$ is the measure provided by the fusion of the features at frame f .

7.1 Prediction

The prediction step relies on the following assumption: if an action state is R_A (resp. F_A) at frame $(f-1)$ then it would be partially R_A (resp. F_A) at frame f . This model of evolution \mathcal{R} (resp. \mathcal{F}) is weighted by a confidence value of $\gamma_{\mathcal{R}} \in [0, 1]$ (resp. $\gamma_{\mathcal{F}} \in [0, 1]$):

Model \mathcal{R} :

If R_A at $(f-1)$ then R_A at f with belief of $\gamma_{\mathcal{R}}$ (10)

Model \mathcal{F} :

If F_A at $(f-1)$ then F_A at f with belief of $\gamma_{\mathcal{F}}$

In the sequel, the following vector notation of a BBA defined on a FoD Ω_A is used:

$$m^{\Omega_A} = [m^{\Omega_A}(\emptyset) \quad m^{\Omega_A}(R_A) \quad m^{\Omega_A}(F_A) \quad m^{\Omega_A}(\Omega_A)]^T$$

A model of evolution can be interpreted as a BBA. For instance, for the model \mathcal{R} :

$$m_{\mathcal{R}}^{\Omega_A} = [0 \quad \gamma_{\mathcal{R}} \quad 0 \quad 1 - \gamma_{\mathcal{R}}]^T \quad (11)$$

The disjunctive rule of combination (Eq. 7) is then used to compute the prediction $\hat{m}_{f, \mathcal{M}}^{\Omega_A}$ from the previous BBA $m_{f-1}^{\Omega_A}$ and the model of evolution $m_{\mathcal{M}}^{\Omega_A}$:

$$\hat{m}_{f, \mathcal{M}}^{\Omega_A} = m_{\mathcal{M}}^{\Omega_A} \odot m_{f-1}^{\Omega_A} \quad (12)$$

The \odot -rule never assigns more belief to an hypothesis than does the previous BBA. As a clue, the prediction with model \mathcal{R} (Eq. 13) is given by:

$$\begin{aligned}
 \hat{m}_{f, \mathcal{R}}^{\Omega_A}(R_A) &= \gamma_{\mathcal{R}} \times m_{f-1}^{\Omega_A}(R_A) \\
 \hat{m}_{f, \mathcal{R}}^{\Omega_A}(\Omega_A) &= (1 - \gamma_{\mathcal{R}}) \times m_{f-1}^{\Omega_A}(R_A) + m_{f-1}^{\Omega_A}(\Omega_A)
 \end{aligned} \quad (13)$$

the others belief are null. When $\gamma_{\mathcal{M}} = 1$, the prediction equals the previous BBA reflecting a total confidence in the current state of action A , while when $\gamma_{\mathcal{M}} = 0$, the model expresses a total ignorance.

7.2 State change

Prediction $\hat{m}_{f,\mathcal{M}}^{\Omega_A}$ and measure $\tilde{m}_f^{\Omega_A}$ represent two distinct pieces of information concerning the state of action A at frame f . They are conjunctively combined (Eq. 7). If the sources are discordant, then a conflict appears indicating a potential state change, i.e. the model might be changed. The conflict value ϵ_f (Eq. 14) is thus relevant for model change requirement:

$$\epsilon_f = (\hat{m}_{f,\mathcal{M}}^{\Omega_A} \odot \tilde{m}_f^{\Omega_A})(\emptyset) \quad (14)$$

The conflict analysis is required to know whether the current model is no longer valid. The CUSUM process of the conflict is well adapted for solving problems concerning *abrupt and short changes* or *gradual and long changes* in the conflict value because it allows to sum up conflict during time.

The initial CUSUM process works as follows: when the CUSUM value becomes greater than a **w**arning threshold \mathcal{T}_w then the frame is stored as f_w and the model is *kept as valid*. As soon as the CUSUM value becomes greater than a **s**top threshold \mathcal{T}_s (at frame f_s) then the model is *changed* and the new model is applied from f_s . When a conflict appears between prediction and measure, as it could be the case in interval $[f_w, f_s]$, it was chosen to *trust the model of evolution*. Thus, the prediction is kept instead of an erroneous measurement and it avoids propagating conflict which is absorptive by the \odot -rule:

$$m_f^{\Omega_A} = \begin{cases} \hat{m}_{f,\mathcal{M}}^{\Omega_A} \odot \tilde{m}_f^{\Omega_A} & \text{if } \epsilon_f = 0 \\ \hat{m}_{f,\mathcal{M}}^{\Omega_A} & \text{otherwise} \end{cases} \quad (15)$$

Eq. (15) accounts for the fact that the BBA $m_{f-1}^{\Omega_A}$ can have only two focal sets (Eq. 13) depending on the current model \mathcal{M} . Furthermore, the output of the TBF is a BBA without conflict and with only one hypothesis whose belief is not null. The interest of the \odot -rule is emphasized when there is often conflict because it allows to obtain $m_{f \rightarrow \infty}^{\Omega_A}(\Omega_A) = 1$ which reflects total ignorance of the system.

To cope with low conflict during a long time, a *fading memory* process has been embedded which allows to forget gradually past event. The fading memory process requires a coefficient nicknamed *fader*, and denoted as λ , which works on the current CUSUM $\mathbf{CS}(f)$ as follows:

$$\mathbf{CS}(f) \leftarrow \mathbf{CS}(f-1) \times \lambda + \epsilon_f \quad (16)$$

The fader is here chosen as a constant and is applied at each frame.

The two models (\mathcal{R} and \mathcal{F}) are tuned once and one model is applied while it is valid. Otherwise, it is changed by the other. A model change is required for an action A_k when the stop threshold \mathcal{T}_w^k is reached by its CUSUM. If the model change is accepted and performed, then the interval of frames $\mathbf{I}_{\mathbf{T}} = [f_w, \min(f_s, f_w + \mathcal{W})]$ can be interpreted as an interval of transition between the two action states. The parameter \mathcal{W} limits the size of the transition. The *vacuous* BBA is assigned to the frames belonging to $\mathbf{I}_{\mathbf{T}}$ to well represent ignorance: $m_{\mathbf{I}_{\mathbf{T}}}^{\Omega_A}(\Omega_A) = 1$. After a model change, the new model is applied from the upper bound of the interval of transition $\mathbf{I}_{\mathbf{T}}$ and the CUSUM is reset.

Remark concerning the initialization procedure: The TBF is an online process. During the initial phase, it is required to determine which is the best model fitting the first data. For that, the CUSUM process is applied on an interval of frames for all models and the chosen one minimizes the CUSUM. The initial TBF output is set to the vacuous belief functions (ignorance, full doubt) with $m_{f_0}^{\Omega_A}(R_A \cup F_A) = 1$ (with f_0 the first frame).

Parameters setting: It is required to set the parameters in a relevant order: first the fader and the models together, then the stop threshold, the warning threshold and at last the window. If the fader λ is too low, then the CUSUM is strongly attenuated. In this case, the stop threshold \mathcal{T}_s has to be small enough to be reachable by the CUSUM. For a given fader λ , the value of the stop threshold \mathcal{T}_s can be estimated if the **s**tart frame f_{sref} is available. Thus, the estimation can be made as follows: the TBF has to be applied with a model of type \mathcal{F} (*false state*) from the beginning of the video sequence and with a stop threshold \mathcal{T}_s unreachable (close to infinity). Then, $\mathcal{T}_s = \mathbf{CS}(f_{sref})$ i.e. the value of the CUSUM at the start frame. If the data do not contain too much conflict, the estimation should be optimal (only one max value and locating at frame f_{sref}) otherwise, the fader has to be increased and the procedure iterated. The mean of the estimated value over a learning set is possible. Concerning the coefficients of the two models ($\gamma_{\mathcal{T}}$ and $\gamma_{\mathcal{F}}$), their role is initially to decrease the conflict between prediction and measure while limiting the variation between two frames.

8 Experiments

The system is tested for action recognition in athletics jumps. The robustness of the system to illumination changes is assessed. At last, a comparison with Bayesian Networks is provided.

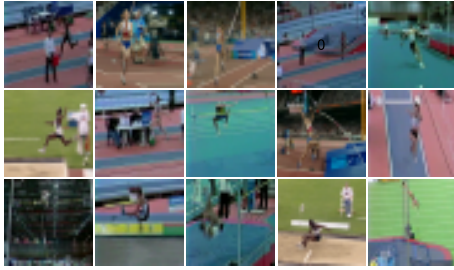


Fig. 9 Running, jumping and falling actions frame which illustrate the diversity of the database.

8.1 Database description

The database used for testing is made of 33 videos acquired with a moving camera and several unknown view angles. The number of frames concerning each action is given in Tab. 4. The database is characterized by its heterogeneity (see Figs. 1 and 9) with a panel of view angles as well as environments and athletes (out/indoor, male, female, other moving people). Indeed, the videos come from TV broadcast recorded either on DVD or VHS tapes (which are digitalized). They are compressed using Divx encoder in 25 fps and 352x288 size. Several meetings are represented such as Olympic Games 2004, French National Championship 2004, Dartfish sequences (<http://www.dartfish.com>) and others samples from TV in 2003. About a half of the database concerns indoor meetings and another half for outdoor. The camera location changes w.r.t. meetings and in each meeting, the camera moves (the camera motion features are automatically estimated) and the view changes (these information are not integrated in the system as a prior knowledge). Since the videos are real, other moving people can appear (Fig. 1) and illumination can change (Figs. 12 and 13). Some of the videos are in slow motion (about 15%). They are all annotated manually (with action true/false labels).

Table 4 Description of the database: running, jumping and falling actions and their corresponding number of frames (cols. 3-5). N_V is the total number of videos.

Jump/Action	N_V	Running	Jumping	Falling	Total
High jump	9	604	351	205	1160
Long jump	8	632	220	213	1065
Pole vault	8	598	417	243	1258
Triple jump	8	676	405	377	1458
Total	33	2510	1393	1038	4941

The test consists in recognizing three actions in four athletics jumps. Actions are: running, jumping and falling. Activities (jumps) are: pole vault, high jump, triple jump and long jump. In addition to camera view variation, other moving people and moving camera, the challenge

of the tests concerns the fact that each video represent one jump and that each jump is made of actions. Therefore, in each video (jump), an action is not separated from the others as usually done in experiments. We assume that the system has to be able to detect actions separately within an activity stream.

A second test is performed with Bayesian Networks in order to compare it with the proposed approach. The test aims at emphasizing the advantage of belief functions and TBM.

8.2 Settings and assessment

The TBF parameters are set once for each action. The setting of the *Temporal Belief Filter* is the same for all actions in pole vault, high jump and long jump (illustrating the robustness of the TBF). In triple jump, the value of the stop threshold for jumping and falling is lower because the duration of these actions in this type of jump is small. The recognition is performed frame by frame and *independently*.

Recall and precision indexes, noted \mathcal{R} and \mathcal{P} respectively, are used for the evaluation and are computed as follows: $\mathcal{R} = \frac{C \cap R}{C}$ and $\mathcal{P} = \frac{C \cap R}{R}$ where C is the reference set obtained by expert annotations, R is the set of retrieved frames provided by the recognition, and $C \cap R$ is the number of correctly retrieved frames.

Since the proposed TBM based system provides belief functions, it is required to take a decision in order to assess it. For that, an action A is considered as true when $m^{\Omega_A}(R_A) > 0$ (this criteria⁹ focuses on the *specific* element R_A , i.e. A is true).

8.3 Recognition performance of the proposed TBM based approach

Tab. 5 gathers the recall and precision indexes for action recognition in each type of jump using the proposed approach based on the Transferable Belief Model (see “TBM” lines). A comparison with usually used Bayesian Networks is also provided (see “BN” lines). The BNs results are discussed further (Section 8.5). The running action is almost the same for each jump accounting for a high overall recognition rate for all jumps. Jumping and falling are well recognized in pole vault and high jump. Results of jumping and falling recognition are less good in the two other types of jumps: in triple jump, these actions have a small duration thus can be deleted by the Temporal Belief Filter, and in long jump, the athlete moves a lot his arms thus disturbing the tracking and therefore the features linked to it. Errors are

⁹ When the BBAs are defined on a FoD with a cardinality greater than 2, the decision must not be taken using belief masses but pignistic probabilities [52].

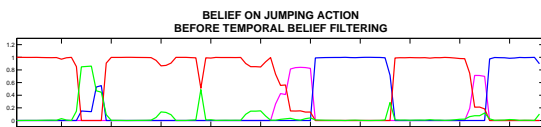
Table 5 Recall \mathcal{R} , precision \mathcal{P} and \mathcal{F}_1 measure (all in %) for the recognition of actions running, jumping and falling by the proposed TBM based method and using Bayesian Networks.

Jump/Action		Running			Jumping			Falling		
		\mathcal{R}	\mathcal{P}	\mathcal{F}_1	\mathcal{R}	\mathcal{P}	\mathcal{F}_1	\mathcal{R}	\mathcal{P}	\mathcal{F}_1
High jump	TBM	97.7	84.4	90.6	74.9	76.4	75.6	75.7	85.2	80.2
	BN	88.7	92.7	90.7	79.9	71.2	75.3	79.3	83.0	81.2
Long jump	TBM	92.0	75.6	83.0	61.2	53.6	57.1	67.1	73.1	70.0
	BN	94.6	70.7	80.9	24.8	57.1	34.6	32.5	72.9	45.0
Pole vault	TBM	85.9	73.2	79.0	77.4	71.2	74.2	75.2	78.6	76.9
	BN	84.0	81.3	82.7	72.1	72.0	72.1	68.8	77.2	72.8
Triple jump	TBM	82.9	64.6	72.6	55.6	66.2	60.4	62.9	53.5	57.8
	BN	91.0	52.6	66.7	33.4	66.9	44.6	50.9	73.7	60.2

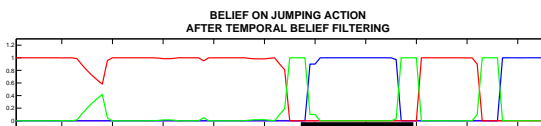
mainly caused by other moving people or objects disrupting tracking. For instance it is the case for the pole in a pole vault because its size is quite important compared to the athlete.

Recognition performance could be improved by a more detailed decomposition of these actions (raising the problem of granularity). Moreover, actions have been described in a static way but dynamic recognition is more relevant. This is challenging because it implies to take into account chaining of events in the TBM framework: this was dealt with a very few times in the past [34,53].

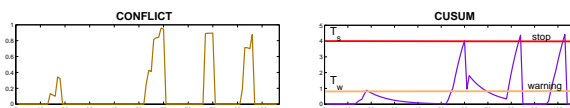
Illustrations of Figs. 10 and 11 concern jumping action in two high jump sequences. Parameters settings are the same for both with: $\lambda = 0.8$, $\mathcal{T}_w = 0.8$, $\mathcal{T}_s = 4$, $\gamma_{\mathcal{R}} = \gamma_{\mathcal{F}} = 0.9$ and $\mathcal{W} = 5$.



(a) Result of the combination of parameters for jumping action recognition. The conflict appears in magenta, the belief on the fact that jumping action is true is in blue, while in red for false and green for imprecise.

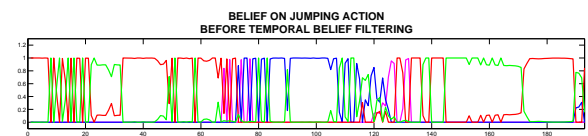


(b) Result after applying the temporal belief filter. The conflict has been converted into ignorance between action states (true and false), and colors of the curves are the same as previously except that the ground truth is underlined in black.

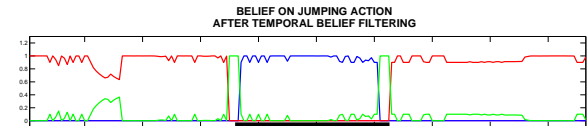


(c) Conflict and CUSUM evolution. Note the memory fading effect.

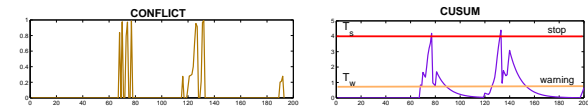
Fig. 10 Result of the high level module: evolution of the belief on the trueness on jumping action in a high jump sequence (normal speed video).



(a) Result of the combination of parameters for jumping action recognition in a high jump. The conflict appears in magenta, the belief on the fact that jumping action is true is in blue, while in red for false and green for imprecise.



(b) Result after applying the temporal belief filter. The conflict has been converted into ignorance between action states (true and false), and colors of the curves are the same as previously except that the ground truth is underlined in black.



(c) Conflict and CUSUM evolution. Note the memory fading effect.

Fig. 11 Result of the high level module: evolution of the belief on the trueness on jumping action in a high jump sequence (slow motion video).

Fig. 10 depicts the impact of the *Temporal Belief Filter* on the recognition (for a normal speed video): Firstly, a *belief specialization* process is automatically performed thus *reducing uncertainty and imprecision* according to the data and models. For instance, from frame 10 to 20, the uncertainty is reduced (green curve). Secondly, *ignorance is expressed in transitions* areas where the conflict is important. From frame 50 to 60, the conflict collapses (magenta curve). *Conflict is thus automatically interpreted as a transition* when located between two different states, e.g. from R_A to F_A . Between frames 95 to 106, a jumping action is detected whereas it is actually a standing up action. This is due to the fact that both camera and human motions (features) in the standing up are close to the ones of a jumping action. In order to distinguish between both actions, a state machine could

be used using the swing value as feature [54]. This “dynamical” recognition using constraints between actions is not the scope of this paper and needs more studies using the TBM. Fig. 11 is a slow motion video sequence. In this particular type of video sequence, the camera motion estimation introduces large discontinuities in related features and so does in the fusion process (Fig. 11(a)). The TBF shows its efficiency for smoothing belief while keeping, even boosting, belief on actions as it can be shown in Fig. 11(b). This high level belief filtering can thus be used instead of usual features filtering.

8.4 Robustness w.r.t. illumination variation

The robustness of the proposed system depends mainly on the camera motion estimation since both the tracking and the recognition processes rely on it. The estimation of the camera motion parameters is based on the minimization of a cost function which embeds the global variation of illumination [39]. Moreover, the software (motion2D) used for the camera motion estimation was already applied in many papers because of its robustness. In this experiment, we show the effect of the illumination variation on the results of this algorithm.

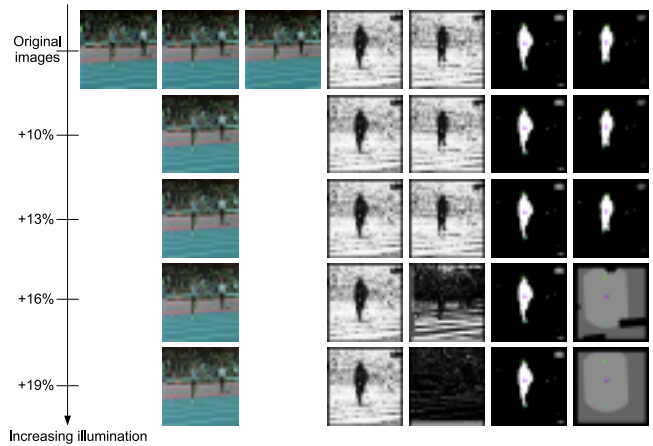
Fig. 12 pictorially describes the impact of the illumination variation in two different high jumps and considering two different environments: one is *indoor* and the other is *outdoor*. Each line of Fig. 12 ((a)-(b)) is made of 7 images:

- three successive input images: one at $f - 1$, f and $f + 1$,
- two dominant motion images (computed by the camera motion estimation): one for the estimation between $f - 1$ and f , and one between f and $f + 1$,
- two images for the tracking results at f and $f + 1$.

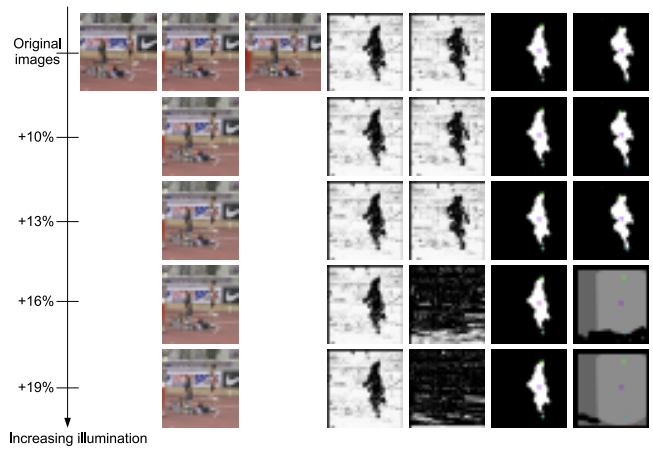
The first line (for Figs. 12(a)-(b)) represents a pattern (reference): the original images (the three first) are such that the global illumination variation is close to 0. Thus the dominant motion images are assumed to be good. Then, for each line (after the first one), the illumination of the second image is artificially increased (by addition of brightness). The value of the variation is given on the left of Fig. 12 ((a)-(b)). The first and third original images are the same as for the first line thus they are not depicted. The estimation algorithm is thus disrupted by a positive variation between $f - 1$ and f and a negative one between f and $f + 1$.

This experiment demonstrates that the limit of variation is of 13%. Beyond this value, the camera motion estimation fails despite the compensation. Obviously, this threshold of 13% is not the same for all images since their characteristics and their own content can disturb the estimation but it gives a rough value for “optimal” conditions.

Fig. 13 pictorially describes the relative variation (in percent) of illumination between two successive images



(a) The indoor case.



(b) The outdoor case.

Fig. 12 Influence of illumination variation on camera motion estimation and tracking for an indoor scene (a) and outdoor (b) in two high jumps. Each line is made of 7 images: three successive input images (at $f - 1$, f and $f + 1$), two dominant motion images (estimation between $f - 1$ and f , and between f and $f + 1$), and two images for the tracking results (at f and $f + 1$). The first line (of (a) and (b)) represents a pattern: the three first images are the original ones with a global illumination variation close to 0. For each line, different from the first, the illumination of the second image is artificially increased (left axis).

in the database for a few videos. Each point of a curve represents the value $100 \times (\bar{I}_f - \bar{I}_{f-1}) / \bar{I}_{f-1}$, with \bar{I}_f the global illumination in frame f . These samples are disturbed cases (considering the whole video). It clearly shows that the illumination variation in the videos remains largely under the limit found previously ($< 13\%$). We can assume that, in the database, the conditions of the application of the camera motion estimation method are fulfilled.

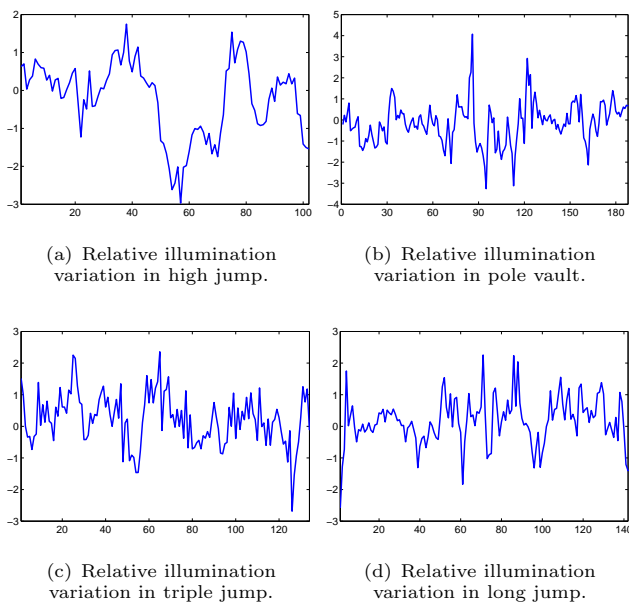


Fig. 13 Each curve depicts the evolution in percent of the relative illumination variation between two successive images in a same jump, one curve corresponds to one video.

8.5 Comparison with Bayesian Networks

The proposed approach is novel because of an original use of the TBM framework as well as the Temporal Belief Filter. Usual methods are based on simple thresholds (precise and certain) or, more frequently, on probabilities which encodes precise and uncertain information. We use belief functions to model imprecise, uncertain and conflicting belief functions.

8.5.1 Methodology The proposed TBM based approach is compared with Bayesian Networks (BNs). BNs machinery is not described in this paper, we just provide the settings. For the tests we have used the Weka software [55]. The set of features (P_{hm} , P_{vm} , P_{div} , P_{veg} , P_{alter} , P_{swing}), described in Tab. 1, is used as input. One BN is used for each type of jump (one for pole vault, one for high jump...). A 10-fold cross validation process is used for the assessment of actions recognition (running, jumping and falling) thus 90% of the dataset is used for learning and 10% for testing. The Minimum Description Length (MDL) criteria is used to learn automatically the topology of the BNs.

8.5.2 Results and analysis The mean of the recognition over the 10 tests is provided in Tab. 5 for each action in each jump. BNs results are less good than the proposed approach based on TBM partly because not sufficient statistics are provided to learn their topology despite of the fact that 90% of the dataset was necessary for learning the complex structure. Bayesian approach is very sensitive to the duration of actions and to the number of images. The same problem appear with HMM.

On the end-user point of view, the TBM based approach is more adapted than BNs. In particular, it is straightforward to add new information and knowledge compared to BNs. This is correlated to the complexity of the topology of BNs: the number of observation symbols is generally large (> 10) and the number of states does not really fit the reality (there is no semantic given to the states). The same problem occurs with HMM [8].

Despite a quite straightforward description of actions and simple but understandable methodology, the proposed TBM based approach leads to good results on this dataset, globally better than Bayesian Networks. One explanation is the representation of doubt, i.e. imprecision, in the belief: only the available is encoded without erroneous *a priori*. Moreover, doubt is modified according to the conflict between features in order to have a coherent fusion process. At last, the proposed methodology is not sensitive to action duration in comparison to BNs.

9 Conclusion and future work

This paper proposes a new architecture for on line human action recognition in athletics sports videos. The novelty of the proposed approach holds in the fact that the Transferable Belief Model framework is used instead of the usual probability theory. The TBM relies on belief functions which are more general than probabilities. The TBM allows to explicitly model and combine the available information, from certain and precise up to total ignorance and emphasizes conflict in the fusion process. The TBM architecture proposed here easily integrates new features or new actions and the description of actions is made understandable for end-users. A Temporal Belief Filter is built in the proposed system in order to make more robust the recognition process and to smooth belief on actions. The proposed architecture is tested on a database composed of 33 athletics videos with moving camera where the purpose is to recognize running, jumping and falling actions in four different types of jumps. Good results were obtained, better than Bayesian Networks despite 90% of the database was used for testing the latter.

In this paper, temporal links between two actions are not integrated: the semantic level only concerns the current state of the behavior i.e. the current action. Thus, the next step of this work is to deal with a sequence of actions corresponding to an activity such as high jump, long jump, triple jump and pole vault. The goal is then to determine at any time if an activity is in progress or finished. This system could be used for video shots classification within a video stream. Its advantage concerns the fact that both actions and activities are recognized. Another fascinating challenge is *adaptation* [8]. Because the recognition process is application dependent, some

knowledge must be *a priori* given by an expert but they are relatively inaccurate and sometimes unreliable. For instance, in HMM, the learning step is very heavy and not adapted for changing environments [8]. We propose, in future work to deal with this problem. In fact, the TBM emphasizes the conflict in the fusion process quantifying inconsistency between sources of information. This information can be used to adapt the models provided by experts to the reality of the data.

10 Acknowledgement

This research is partially supported by SIMILAR European excellence network. The authors thank the Vista research team at Irisa/Inria Rennes (France) for the use of the Motion2D software.

References

1. W. Hu, T. Tan, L. Wang, and S. Maybank. A survey on visual surveillance of object motion and behaviors. *IEEE Trans. on Systems, man and cybernetics C*, 34(3), 2004.
2. K. Messer, W.J. Christmas, E. Jaser, J. Kittler, B. Levenaise-Obadia, and D. Koubaroulis. A unified approach to the generation of semantic cues for sports video annotation. *Signal Processing*, 85:357–383, 2005.
3. A. Jaimes and N. Sebe. Multimodal human computer interaction: A survey. In *IEEE Int. Workshop on Human Computer Interaction in conjunction with ICCV*, volume 3766, pages 1–15, Beijing, China, 2005.
4. B. Li, J.H. Errico, H. Pan, and I. Sezan. Bridging the semantic gap in sports video retrieval and summarization. *Jour. of Visual Communication and Image Representation*, 15, 2004.
5. D.A. Sadlier and N.E. O'Connor. Event detection in field sports video using audio-visual features and a support vector machine. *IEEE Trans. on Circuits and Systems for Video Technology*, 15(10), 2005.
6. M. Lew, N. Sebe, and J. Eakins. Challenges in image and video retrieval. *Lecture notes in Computer Science, ICIVR*, 2383:1–6, 2002.
7. N.D. Freitas, E. Brochu, K. Barnard, P. Duygulu, and D. Forsyth. Bayesian models for massive multimedia databases: A new frontier. In *Valencia Int. Meeting on Bayesian Statistics/2002 ISBA Int. Meeting*, 2002.
8. M. Shah. Understanding human behavior from motion imagery. *Machine Vision and Applications*, 14:210–214, 2003.
9. S. Hongeng, R. Nevatia, and F. Bremond. Video-based event recognition and probabilistic recognition methods. *Computer Vision and Image Understanding*, 96:129–162, 2004.
10. G.J. Klir and M.J. Wierman. *Uncertainty-based information. Elements of generalized information theory, 2nd edition*. Studies in fuzzyness and soft computing. Physica-Verlag, 1999.
11. D. Dubois, M. Grabisch, H. Prade, and Ph. Smets. Using the transferable belief model and a qualitative possibility theory approach on an illustrative example: the assessment of the value of a candidate. *Int. Jour. of Intelligent Systems*, 16:1245–1272, 2001.
12. B. Ristic and P. Smets. Target identification using belief functions and implication rules. *IEEE Trans. Aerospace and Electronic Systems*, 41(3):1097–1102, 2005.
13. M. Petkovic and W. Jonker. Integrated use of different content derivation techniques within a multimedia database management system. *Jour. of Visual Communication and Image Representation*, 15:303–329, 2004.
14. P. Smets and R. Kennes. The Transferable Belief Model. *Artificial Intelligence*, 66(2):191–234, 1994.
15. G. Shafer. *A mathematical theory of Evidence*. Princeton University Press, Princeton, NJ, 1976.
16. P. Smets. *Advances in the Dempster-Shafer Theory of Evidence - What is Dempster-Shafer's model ?*, pages 5–34. Wiley, R.R. Yager and M. Fedrizzi and J. Kacprzyk edition, 1994.
17. E. Ramasso, M. Rombaut, and D. Pellerin. A Temporal Belief Filter improving human action recognition in videos. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 141–144, 2006.
18. J. Schubert. Clustering belief functions based on attracting and conflicting metalevel evidence using Potts spin mean field theory. *Information Fusion*, 5(4):309–318, 2004.
19. L. Wang, W. Hu, and T. Tan. Recent developments in human motion analysis. *Pattern Recognition*, 36(3):585–601, 2003.
20. Y. Yacoub and M. Black. Parametrized modeling and recognition of activities. *Computer Vision and Image Understanding*, 73(2):232–247, 1999.
21. L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. of the IEEE*, 77:257–285, 1989.
22. K. P. Murphy. *Dynamic Bayesian Networks: Representation, inference and learning*. PhD thesis, UC Berkeley (CSD), 2002.
23. T. Xiang and S. Gong. Discovering Bayesian causality among visual events in a complex outdoor scene. In *IEEE on Advanced Video and Signal based Surveillance*, pages 177–182, 2003.
24. Y. Luo, T.D. Wu, and J.N. Hwang. Object-based analysis and interpretation of human motion in sports video sequences by dynamic Bayesian networks. *Computer Vision and Image Understanding*, 92:196–216, 2003.
25. L. Vefghi and D.A. Linkens. Dynamic monitoring and control of patient anaesthetic and dose levels: Time-Delay, Moving-Average Neural Networks, and Principal Components Analysis. *Computer Methods and Programs in Biomedicine*, 59:91–106, 1999.
26. D. Zhong and S.-F. Chang. Real-time view recognition and event detection for sports video. *Jour. of Visual Communication and Image Representation*, 15:330–347, 2004.
27. G. Medioni, I. Cohen, F. Brémond, S. Hongeng, and R. Nevatia. Event detection and analysis from video streams. *PAMI*, 23(8):873–889, 2001.
28. D. Ayers and M. Shah. Monitoring human behavior from video taken in an office environment. *Image and Vision Computing*, 13:833–846, 2001.

29. Z. Ding, H. Bunke, M. Schneider, and A. Kandel. Fuzzy timed Petri net : Definitions, properties, and applications. *Mathematical and Computer Modelling*, 41:345–360, 2005.
30. S.J. Lee and P.H. Seong. Development of automated operating procedure system using fuzzy colored Petri nets for nuclear power plants. *Annals of nuclear energy*, 31:849–869, 2004.
31. A. Fay. A fuzzy knowledge-based system for railway traffic control. *Engineering applications of Artificial Intelligence*, 13:719–729, 2000.
32. N. Bourbakis, J.R. Gattiker, and G. Bebis. Interpreting a dynamic and uncertain world: Task-based control. *Int. Jour. of Artificial Intelligence Tools*, 12(1):5–85, 2003.
33. J.M. Nigro, S. Loriette-Rougegrez, and M. Rombaut. Driving situation recognition with uncertainty management and rule-based systems. *Engineering Applications of Artificial Intelligence*, 15:217–228, 2002.
34. M. Rombaut, I. Jarkass, and T. Denoeux. State recognition in discret dynamical systems using Petri nets and Evidence theory. In *Europ. Conf. on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, 1999.
35. V. Girondel, A. Caplier, L. Bonnaud, and M. Rombaut. Belief theory-based classifiers comparison for static human body postures recognition in video. *Int. Jour. of Signal Processing*, 2(1):29–33, 2005.
36. Z. Hammal, A. Caplier, and M. Rombaut. Belief theory applied to facial expressions classification. In *Int. Conf. on Advances in Pattern Recognition*, Bath, United Kingdom, 2005.
37. T.B. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 81:231–268, 2001.
38. J. Wang and S. Singh. Video analysis of human dynamics-a survey. *Real-Time Imaging*, 9(5):321–346, 2003.
39. J.M. Odobez and P. Bouthemy. Robust multiresolution estimation of parametric motion models. *Jour. of Visual Communication and Image Representation*, 6(4):348–365, 1995.
40. G. Piriou, P. Bouthemy, N. Peyrard, and J.F. Yao. Probabilistic models of image motion for recognition of dynamic content in video. In *Int. Workshop on Computer Vision and Image Analysis*, volume 11, Las Palmas de Gran Canaria, Spain, 2002.
41. R. Fablet, P. Bouthemy, and P. Perez. Non parametric motion characterization using causal probabilistic models for video indexing and retrieval. *IEEE Trans. on Image Processing*, 11(4):393–407, 2002.
42. C. Panagiotakis and G. Tziritas. Recognition and tracking of the members of a moving human body. In *Articulated Motion and Deformable Objects*, pages 86–98, 2004.
43. B. Yaghlane, P. Smets, and K. Mellouli. Independence concept for belief functions. In *Technologies for constructing intelligent systems: tools*, Heidelberg, Germany, 2002. Physica-Verlag GmbH.
44. A. Ayoun and Ph. Smets. Data association in multi-target detection using the transferable belief model, 2000.
45. M. Zribi and M. Benjelloun. Parametric estimation of Dempster-Shafer belief functions. In *Int. Conf. on Information Fusion*, pages 485–491, 2003.
46. Z. Elouedi, K. Mellouli, and Ph. Smets. Assessing sensor reliability for multisensor data fusion within the transferable belief model. *IEEE Trans. Systems, Man and Cybernetics*, 34(1):782–787, 2004.
47. P. Smets. Beliefs functions: the disjunctive rule of combination and the Generalized Bayesian Theorem. *Int. Jour. of Approximate Reasoning*, 9:1–35, 1993.
48. M. Rombaut and Y.M. Zhu. Study of Dempster-Shafer theory for image segmentation applications. *Image and Vision Computing*, 20(1):15–23, 2002.
49. T. Denoeux and A.B. Yaghlane. Approximating the combination of belief functions using the fast moebius transform in a coarsened frame. *Int. Jour. of Approximate Reasoning*, 37:77–101, 2002.
50. Philippe Smets. The application of the matrix calculus to belief functions. *Int. Jour. of Approximate Reasoning*, 31(1-2):1–30, October 2002.
51. R. Haenni and N. Lehmann. Implementing belief function computations. *Int. Jour. of Intelligent Systems*, 18(1):31–49, 2003.
52. Ph. Smets. Decision making in the TBM: the necessity of the pignistic transformation. *Int. Jour. of Approximate Reasoning*, 38:133–147, 2005.
53. E. Ramasso, D. Pellerin, and M. Rombaut. Belief Scheduling for the recognition of human action sequence. In *Int. Conf. on Information Fusion*, pages 1–8, Florence, Italia, 2006.
54. C. Panagiotakis, E. Ramasso, G. Tziritas, M. Rombaut, and D. Pellerin. Shape-motion based athlete tracking for multilevel action recognition. In F.J. Perales and R.B. Fisher, editors, *Proc. of the 4th Int. Conf. on Articulated Motion and Deformable Objects*, pages 385–394. Springer-Verlag, 2006.
55. I.H. Witten and E. Frank. *Data-mining: practical machine learning tools and techniques*. 2nd edition. Morgan Kaufman, 2005.