Shape-Motion Based Athlete Tracking for Multilevel Action Recognition

Costas Panagiotakis[1], Emmanuel Ramasso[2], Georgios Tziritas[1],
Michèle Rombaut[2], and Denis Pellerin[2]

[1] Department of Computer Science, University of Crete, P.O. Box 2208,
Heraklion, Greece
{cpanag, tziritas}@csd.uoc.gr
[2] Laboratoire des Images et des Signaux, 46 avenue Félix Viallet,
38031 Grenoble, France
first_name.family_name@lis.inpg.fr

Abstract. An automatic human shape-motion analysis method based on a fusion architecture is proposed for human action recognition in videos. Robust shape-motion features are extracted from human points detection and tracking. The features are combined within the Transferable Belief Model (TBM) framework for action recognition. The TBM-based modelling and fusion process allows to take into account imprecision, uncertainty and conflict inherent to the features. Action recognition is performed by a multilevel analysis. The sequencing is exploited for feedback information extraction in order to improve tracking results. The system is tested on real videos of athletics meetings to recognize four types of jumps: high jump, pole vault, triple jump and long jump.

1 Introduction

Human motion analysis has many applications in many areas, such as analysis of athletic events, surveillance, content-based image storage and retrieval. The main scientific challenges in human motion analysis are to detect, track and identify people and to recognize the human activity [1] from observations coming from video. Wang, Hu and Tan [2] emphasize on three major issues of human motion analysis systems, namely human detection, tracking and activity understanding. There are model based approaches and systems using Shape-From-Silhouette methods to detect and track the human in 2D [3]. The silhouettes are generally of good quality providing valuable information about the position and shape of the person. Camera motion estimation methods [4] can locate the independently moving objects.

Many methods have been proposed for action recognition [2] notably based on *classification*, *template matching* and *neural networks*. Generally, the methods are based on the *Bayesian framework* with *Hidden Markov Models* (HMM) and *Dynamic Bayesian Network* (DBN) [5]. Other methods are developed in Artificial Intelligence community notably *Petri Nets* [6]. In [7], it is proposed an

F.J. Perales and R.B. Fisher (Eds.): AMDO 2006, LNCS 4069, pp. 385–394, 2006.

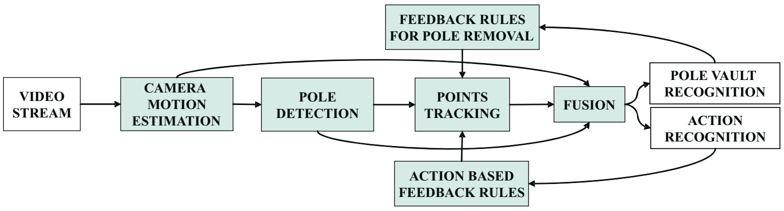


Fig. 1. Schema of the proposed system architecture

architecture for human action recognition using the *Transferable Belief Model* (TBM) which is based on belief theory.

A challenging problem appears when the camera is moving and the estimated human silhouettes are of low quality or extremely wrong (see Fig. 4(a)). In this work we focus on automatic human detection, tracking and action recognition under real and dynamic environments of athletic meetings. We suppose that the camera tracks the athlete and we test the algorithm in sports such as pole vault, high jump, triple jump and long jump.

The proposed architecture consists of several main modules (Fig. 1):

1. Silhouettes are computed using a camera motion estimation method [4], where an affine model is used to describe the camera motion. Such a model is generally sufficient for most of real video sequences. The above method that we use, was implemented by the Vista Team of IRISA.
2. The pole detection procedure, is applied to the human silhouette detecting the pole and extracting features related to it such as its eccentricity and its position.
3. Four major human points are recognized and tracked using the human silhouettes. Shape-motion based features are extracted using the results of the tracking procedure.
4. A fusion architecture, based on TBM, is used for action recognition. The input parameters for the fusion process include camera motion, pole detection and human shape-motion parameters estimated by the corresponding modules.
5. The results of the fusion process can be used as feedback information improving the results of human tracking.

The rest of the paper is organized as follows: Section 2 presents the human shape-motion analysis method. Section 3 describes the action recognition and feedback method. Finally, Sections 4 and 5 provide experimental results and the discussion, respectively.

2 Human Shape-Motion Analysis

The human shape-motion analysis is based on binary silhouettes. They are computed from camera motion estimation as described in [7].

2.1 Pole Detection

The pole is recognized first since it can be easily detected by its shape which has high eccentricity. The eccentricity (ε) is defined by the ratio between the two principal axes of the best fit ellipse, measuring how thin and long a region is. If the detected region has high ε (more than 20) then it is probably a pole. This feature is relevant in the fusion process to recognize the pole vault videos.

First, the highest area object (O_1) is detected. Then, the end of pole point (P_e) is estimated. P_e is defined as the farthest O_1 point from the mass center (C) of O_1 object under the constraint that it is found above the C as the athlete is running. The pole pixels will be detected by a region growing method (RG) starting from P_e point. This method terminates when the area of region exceeds the 50% of the O_1 area or when the number of pixels of the boundary between the region and O_1 exceeds a threshold. The threshold is a percentage (e.g. 40%) of the square root of the O_1 area approximating the double of O_1 mean width. However, the region will have been expanded in the athlete area. Therefore, we have to ignore the last pixels that RG adds, until the region where ε will be maximum (see Fig. 2). Let O_2 be the estimated pole region. We compute the distance d between the farthest point (P_f) of O_2 from P_e and P_e itself. Then, ε can be estimated by the ratio $\varepsilon = \frac{\pi d^2}{O_{2\ area}}$. P_f can be approximated directly by the last point that the RG method adds.

The proposed pole detection method detects the pole with high accuracy and robustness to silhouette noise (see Fig. 2(e)). The strong point of this method is that it is simple and low cost. The results on our database show a great performance of this detector.

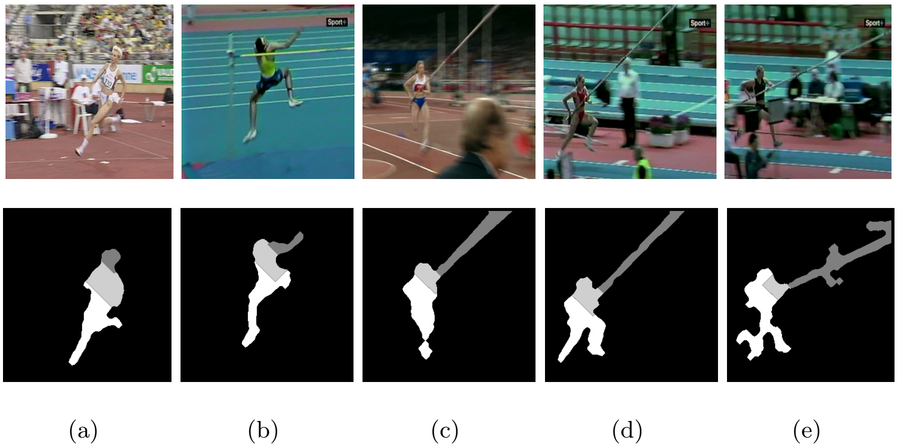


Fig. 2. Results of pole detection procedure. The light gray pixels denote those that ignored (last added) by the RG method and the gray pixels denote the detected pole region. **(a)** $\varepsilon = 6.08$, **(b)** $\varepsilon = 12.24$, **(c)** $\varepsilon = 31.27$ **(d)** $\varepsilon = 50.01$, **(e)** $\varepsilon = 31.32$.

2.2 Points Detection and Tracking

In this step, four major human points, namely: the head center, the mass center, the left end of leg and the right end of leg (see Fig. 4(b)) are detected and tracked using as input human silhouettes. The above points are selected because they are visible in the whole sequence providing sufficient information for the action recognition. The method is divided into two procedures: the detection procedure and the tracking procedure. Results of this method are illustrated in Fig. 3.

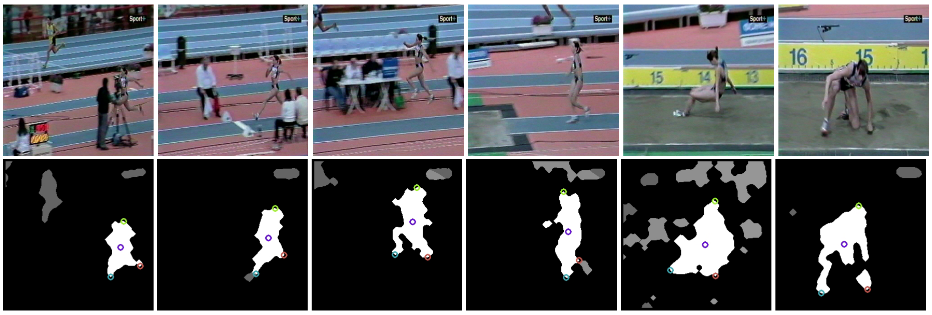


Fig. 3. Results of Major Human Points Tracking method on triple jump sequence

Detection. In this step, the four major human points are automatically detected (see Fig. 4(b)). This procedure is executed just once, in the first silhouette frame of the sequence or when the tracking history is erased by feedback information of the fusion process. The "Human Points Detection" algorithm is described hereafter.

First, the mass center point (C) is computed. This point is defined as the mass center of the foreground pixels. Next, the human body major axis (see Fig. 4(b)) is computed using second order moments. The head point (H) is defined as the farthest major axis point from C, that is found above the C. The first end of leg point (L_1) can be computed by getting the farthest foreground pixel from the C, that is found below the C. Finally, the next end of leg point (L_2) should have the following properties: high distances from C, H and L_1. Moreover, the triangle PCL_1 should be close to an isosceles triangle, where P denotes a candidate L_2 point. The last two constraints are equal to the triangle area ($E(PCL_1)$) maximization. Thus, the maximization of product ($|PH| \cdot |PC| \cdot E(PCL_1)$) provides the L_2 point.

Tracking. In this step, the four major human points are tracked. This procedure is executed in every frame of the sequence, apart from the first one, taking as input the position of the four major human points in the previous frame (history) and the current silhouette image.

First, we reclassify the binary silhouette image pixels reducing the number of wrong classified pixels. We compute the minimum distance of each foreground object from the previous position of the four human points multiplied by the percentage of the foreground pixels that belong to a line segment started on the mass center of the foreground object and ended on the specific major human point. If this distance is higher than a threshold then the foreground pixels will be classified to background class (gray pixels of Fig. 4(b)).

The four major human points can be detected by "Human Points Detection". This method produces two pairs of solutions for the head point and the leg points, 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.

2.3 Human Shape-Motion Parameters

Using the results of pole detection and points tracking, we can compute shape-motion features useful for action recognition. The estimated pole eccentricity (ε) is relevant shape feature since we can recognize if the detected region is a pole. It can also be used to detect dropping bar during jumping or falling stages in high jump and pole vault.

The motion based features are computed from the major points trajectories. One important feature concerns the vertical translation of the mass center (P_{msvt}). Then, the angle between the human major axis and the horizontal axis (Θ_1) (see Fig. 4(c)) is of key of importance for action discrimination. If this angle is about 90^o, the human is standing or running, whereas important variation occur during the jumping and falling in high jump and pole vault. Moreover, the angle between the legs (Θ_{34}) (see Fig. 4(c)) is another relevant feature. Indeed, the gait period can be measured from its trajectory providing an estimation of the human speed. The camera motion parameters are also exploited for action recognition: the camera horizontal translation (P_{cht}), the camera vertical translation (P_{cvt}), and the camera zoom (P_{cz}).

3 Human Action Sequence Recognition

The parameters described previously are now combined within TBM [8] framework for action recognition. Some parts of the work described in the sequel relies on [7,9].

3.1 From Numerical Parameters to Belief on Actions

An action A is described by two states gathered in the frame of discernment (FoD) $\Omega_A = \{R_A, F_A\}$ with R_A (resp. F_A) stands for "action A is right" (resp. "A is *false*"). A basic belief assignment (BBA) on an A according to a parameter P is defined on the set of propositions $2^{\Omega_A} = \{\emptyset, R_A, F_A, R_A \cup F_A\}$ by $m_P^{\Omega_A} : 2^{\Omega_A} \to [0,1]$, $X \to m_P^{\Omega_A}(X)$ and by construction $m_P^{\Omega_A}(\emptyset) = 0$, and $\sum_{X \subseteq \Omega_A} m_P^{\Omega_A}(X) = 1$. The set $R_A \cup F_A$ explicitly represents the doubt concerning the real state of an action: it does not imply any additional claims regarding the subsets, i.e. neither R_A nor F_A. This is a fundamental difference with a probability measure which is additive. A fuzzy-set inspired method [7] is used

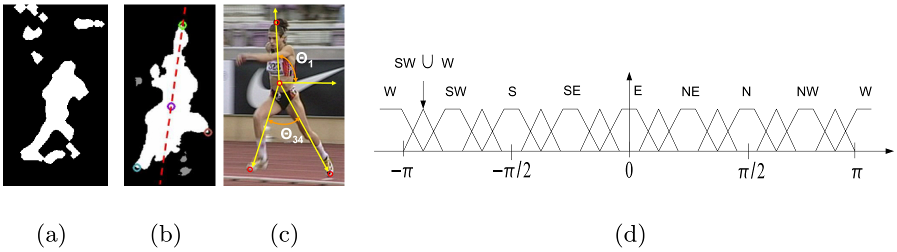


Fig. 4. (a) Low quality silhouette. **(b)** Estimated four major human points. The human body major axis is shown as a red dashed line. **(c)** The human major axis angle (Θ_1) and the angle between legs (Θ_{34}). **(d)** Numeric-to-symbolic conversion of Θ_1.

to convert each numerical parameter described section 2.3 into sources of belief (see Fig. 4(d)).

3.2 Transferable Belief Model Fusion

Belief of several parameters are combined in the axiomatically well-founded Transferable Belief Model (TBM) framework proposed by Smets and Kennes [8] to obtain a belief which takes all parameters into account. The fusion process is performed frame by frame for each action independently by rules of combination defined for two distinct BBAs $m_{P_1}^{\Omega_A}$ and $m_{P_2}^{\Omega_A}$ by:

$$m_{P_1}^{\Omega_A} \textcircled{\triangle} m_{P_2}^{\Omega_A}(E) = \sum_{C \triangle D = E} m_{P_1}^{\Omega_A}(C).m_{P_2}^{\Omega_A}(D) \qquad (1)$$

with $\triangle = \cap$ (resp. $\cup$) for the conjunctive (resp. disjunctive) rule of combination. The rules of combination can be used in logical rules such as "*if* ... AND ... OR ... *then* ..." for describing actions by means of parameters states. These logical rules are then translated into belief combinations where the logical AND is replaced by the $\textcircled{\cap}$-rule and the logical OR by the $\textcircled{\cup}$-rule assuming the same FoD [8]. Some reliability factors can also be integrated in equation (1).

3.3 From Action to Sequence of Actions

The Temporal Belief Filter (TBF) proposed in [9] is exploited for *action sequence recognition*. The TBF worked on each action independently taking as input the BBA obtained from parameters fusion and providing a temporally clean and consistent BBA.

The TBF dissociates in an online manner the intervals of frames where an action is *right* to the intervals of frames where the action is *false*. For that, the current state is predicted and conjunctively combined with the measurements resulting in a smooth belief. The state change detection is based on the conflict between prediction and measurements and computed by the conjunctive rule of combination. The state change detector embeds a CUSUM process of the conflict to be more robust. While the CUSUM process does not indicate that the state has to be changed, the state is compelled even if there is conflict between prediction and measurements accounting for a smooth belief.

We assume a sequence $S_n = \{A_1^n \rightarrow A_2^n \rightarrow \ldots \rightarrow A_k^n \rightarrow \ldots \rightarrow A_K^n\}$ made of K actions. The sequences evolutes from an action $\{A_k^n\}$ to $\{A_{k+1}^n\}$ if the TBF indicates that $\{A_k^n\}$ becomes *false* or if $\{A_{k+1}^n\}$ becomes *right*. The action sequencing method ensures that, at each frame of the video, one and only one action is in the *right state* while the others are in the *false state*. The final goal of action sequencing is to find out which sequence better matches the data at each frame of the video. For that, a Quality Performance Criteria (QRP) is proposed.

When the sequence S_n evolutes from $\{A_k^n\}$ to $\{A_{k+1}^n\}$, a Local QRP ($\mathbf{LQRP}_k^n$) is computed for $\{A_k^n\}$. This criterion is computed without reference for a given action thus it is "local" w.r.t the sequence. The $\mathbf{LQRP}_k^n$ is defined by the mean of pignistic probability [8] of action $\{A_k^n\}$ weighted by the contradiction[1] between the data and the state compelled by the TBF. When the entire sequence is covered, K values of $\mathbf{LQRP}_k^n$ are available. A Global QRP ($\mathbf{GQRP}^n$) is computed by the mean of the $\mathbf{LQRP}_k^n$: $\mathbf{GQRP}^n = \sum_{k=1}^{K} \mathbf{LQRP}_k^n / K$. The sequence S_n better corresponds to the data than S_p if $\mathbf{GQRP}^n > \mathbf{GQRP}^p$ and if $\mathbf{GQRP}^n$ is greater than a given required value (e.g. 50%).

3.4 Coarse to Fine Approach and Feedback

The action sequence method consists in two steps: a coarse detection and a fine detection of the actions. The coarse step involves the camera motion parameters and the center of mass. In the fine step, sequencing based on Θ_1 is used to discriminate all actions.

Coarse step. The sequences to recognize concern four types of jump: high jump (S_{hj}), pole vault (S_{pv}), triple jump (S_{tj}) and long jump (S_{lj}). Sequences S_n, $\forall n \in \{hj, pv, lj\}$ are firstly described by a *coarse* action sequence: $S_n = \{R_n \rightarrow J_n \rightarrow F_n \rightarrow U_n\}$, where $\{R_n\}$ is the running action, $\{J_n\}$ is jumping, $\{F_n\}$ is falling and $\{U_n\}$ is standing up in sequence S_n. For triple jump, the coarse sequence is: $S_{tj} = \{R_{tj} \rightarrow J_{tj} \rightarrow F_{tj} \rightarrow J_{tj} \rightarrow F_{tj} \rightarrow J_{tj} \rightarrow F_{tj} \rightarrow U_{tj}\}$. There is no subsequence for triple jump because the coarse one is characteristic and can not be confused with the other types of jump.

All actions $\{R_n, J_n, F_n, U_n\}, \forall n \in \{hj, pv, lj, tj\}$ are detected by a fusion process performed at each frame of the video following these rules (see Section 2.3 for symbols):

IF ($\mathrm{P_{cht}}$ is high **OR** $\mathrm{P_z}$ is high **OR** $\mathrm{P_{msvt}}$ is almost null)
THEN ($\{\mathrm{R_n}\}$ is true)

IF ($\mathrm{P_{cvt}}$ is highly positive **OR** $\mathrm{P_{msvt}}$ is highly positive)
THEN ($\{\mathrm{J_n}\}$ and $\{\mathrm{U_n}\}$ are true)

IF ($\mathrm{P_{cvt}}$ is highly negative **OR** $\mathrm{P_{msvt}}$ is highly negative)
THEN ($\{\mathrm{F_n}\}$ is true)

[1] This information is provided by the TBF, see equation (9) of [9].

Rules are well-managed in the TBM using eq. (1). The coarse definition of a sequence provides the intervals of frame where an action is potentially true but does not allows to distinguish the type of sequence. In order to differentiate the sequences, a fine analysis is required.

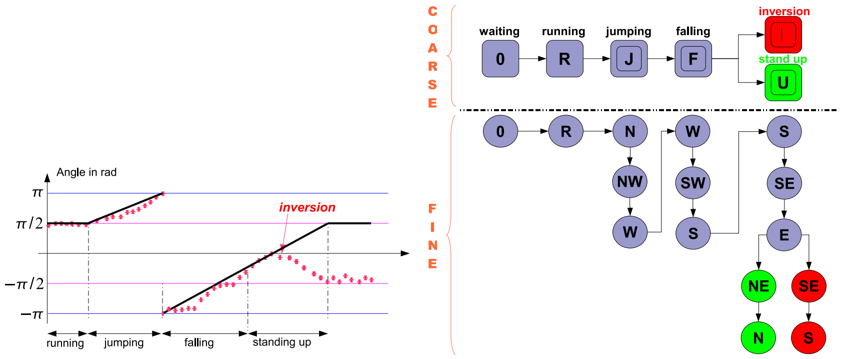
Fine step. The fine analysis is performed in the intervals of frame detected by the coarse process by exploiting the parameter Θ_1. The numerical-to-symbolic conversion [7] of Θ_1 is performed by dividing the interval of possible values $[-180^o, 180^o]$ into 4 main positions $\{N, S, W, E\}$ (North, South, West, East) and 4 intermediate positions $\{NW, SW, SE, NE\}$. The conversion is depicted Fig. 4(d) and shows the explicit modelling of the doubt between two positions, for instance $SW \cup W$. The fuzzy description of the angle value allows to take imprecision and uncertainty of this parameter into account. Notably, each position is modelled by a trapezoidal fuzzy set with a size support of 40^o.

The sequencing of the angle value is performed according to each action sequence. One set of sequences is necessary for both right-to-left and left-to-right translations of the camera. In Table 1, only the first case is described. In Fig. 5(b), the high jump action sequence is pictorially described.

Table 1. Sequences of the angle for each type of jump

sequence name	symbol and action sequence expression
pole vault	$S_{pv} = \{R_{pv} \rightarrow J_{pv} \rightarrow F_{pv} \rightarrow U_{pv}\}$
running	$R_{pv} = \{N \cup (\varepsilon\ is\ high)\}$
jumping	$J_{pv} = \{N \rightarrow NE \rightarrow E \rightarrow SE \rightarrow S \rightarrow SE \rightarrow E\}$
falling	$F_{pv} = \{E \rightarrow NE \rightarrow N \rightarrow NW \rightarrow W\}$
standing up	$U_{pv} = \{W \rightarrow NW \rightarrow N\}$
high jump	$S_{hj} = \{R_{hj} \rightarrow J_{hj} \rightarrow F_{hj} \rightarrow U_{hj}\}$
running	$R_{hj} = \{N\}$
jumping	$J_{hj} = \{N \rightarrow NW \rightarrow W\}$
falling	$F_{hj} = \{W \rightarrow SW \rightarrow S\}$
standing up	$U_{hj} = \{S \rightarrow SE \rightarrow E \rightarrow NE \rightarrow N\}$
long jump	$S_{lj} = \{R_{lj} \rightarrow J_{lj} \rightarrow F_{lj} \rightarrow U_{lj}\}$
running	$R_{lj} = \{N\}$
jumping	$J_{lj} = \{N\}$
falling	$F_{lj} = \{N \rightarrow NE \rightarrow E\}$
standing up	$U_{lj} = \{E \rightarrow NE \rightarrow N\}$

Error detection for feedback. A feedback is a powerful means to adapt a processing chain to varying conditions. In order to illustrate the approach, the example of high jump is presented. In Fig. 5(a), the angle shows an inversion of the human points provided by the tracking due to very bad segmentation when the athlete falls on the air mattress (top foot, down head). This error can be detected by means of action sequencing (Fig. 5(b)). We denote I_{hj} the symbol of the action associated to the inversion in a high jump. Coarsely, the inversion is searched after a falling. Finely, the sequence used to detect this error is close to the sequence used for a standing up: $I_{hj}^{\Theta_1} = \{S, SE, E, SE, E\}$. This sequence is depicted in Figs.5(a) and 5(b). When the error sequence is of high quality,



(a) (b)

Fig. 5. (a) Theoretical angle rough evolution (full line) and observed one (dotted-line). **(b)** Action sequence by a coarse to fine approach for high jump based on angle Θ_1.

i.e. **GQRP** is high, then an error is assumed to be detected and a feedback process is performed onto the tracking algorithm to correct the inversion. The same reasoning can be applied for others jumps, notably for pole vault.

4 Experiments

The database contains 68 videos with four types of jumps: high jump (hj), pole vault (pv), triple jump (tj) and long jump (lj). Each video is analyzed by the four sequences S_n, $\forall n \in \{hj, pv, lj, tj\}$ providing four criteria $\mathbf{GQRP}^n$. A jump n^* is associated to the current video if $n^* = \max_n \mathbf{GQRP}^n$ (Section 3.3) and if $\mathbf{GQRP}^{n^*}$ is greater than 50%. One setting per type of jump is provided for the TBF. Then, the obtained results are compared with the manually annotated video to compute a precision index. Using the coarse sequencing, all actions are well detected. However, to discriminate actions, we use the refinement described Section 3.4 and based on the angle.

The *error rates* are: $\mathbf{E}_{hj} = 2/15$, $\mathbf{E}_{pv} = 4/26$, $\mathbf{E}_{tj} = 3/12$ and $\mathbf{E}_{lj} = 4/15$. Concerning *inversion of the tracked points in high jump*, the detection rate is of $\mathbf{C}_{inv-hj} = 6/8$. The reasons have been identified to account for error rates: videos with pure divergence (zoom) with athlete in front of the camera prevent from using the angle, bad pole deletion, video shot changes and bad camera motion estimation in too low quality videos disturb the tracking.

5 Conclusion

An unsupervised-automatic human motion analysis and action sequence recognition (running, jumping and falling, standing up) based on the TBM is proposed and tested on athletics videos. The first main contribution concerns the original robust human shape-motion parameters extractors from camera motion and human silhouette. The color independent silhouette analysis algorithm detects and tracks four major human points. Sometimes, the tracking procedure fails because of wrong previous silhouettes (wrong history) or because of pole appearing in pole vault sequences (wrong shape). We have developed a shape based pole detector, detecting automatically the pole vault videos and removing the pole with great pole detection ratio. The second main contribution concerns the action sequence recognition based on a fusion process using the TBM. A multilevel approach is exploited to refine action detection and recognition. Some action sequences are also used to detect errors in tracking providing feedback information for further corrections.

Acknowledgments

This work is partially supported by SIMILAR European Network of Excellence.

References

1. J. Aggarwal and S. Park, "Human motion: Modeling and recognition of actions and interactions," in *3DPVT04*, 2004, pp. 640–647.
2. L. Wang, W. Hu, and T. Tan, "Recent developments in human motion analysis," *PR*, vol. 36, no. 3, pp. 585–601, 2003.
3. C. Panagiotakis and G. Tziritas, "Recognition and tracking of the members of a moving human body," in *Proc. of AMDO 2004*, 2004, pp. 86–98.
4. J.M. Odobez and P. Bouthemy, "Robust multiresolution estimation of parametric motion models," *J. of Vis. Comm. and Image R.*, vol. 6, no. 4, pp. 348–365, 1995.
5. 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," *CVIU*, vol. 92, pp. 196–216, 2003.
6. M. Rombaut, I. Jarkass, and T. Denoeux, "State recognition in discrete dynamical systems using petri nets and evidence theory," in *ECSQARU*, June 1999.
7. E. Ramasso, D. Pellerin, C. Panagiotakis, M. Rombaut, G. Tziritas, and W. Lim, "Spatio-temporal information fusion for human action recognition in videos," in *13th European Signal Processing Conf.*, 2005.
8. P. Smets and R. Kennes, "The Transferable Belief Model," *Artificial Intelligence*, vol. 66, no. 2, pp. 191–234, 1994.
9. E. Ramasso, M. Rombaut, and D. Pellerin, "A temporal belief filter improving human action recognition in videos," in *ICASSP*, 2006, to appear.