SPATIOTEMPORAL INFORMATION FUSION FOR HUMAN ACTION RECOGNITION IN VIDEOS

Emmanuel Ramasso[†], *Denis Pellerin*[†], *Costas Panagiotakis*[‡], *Michèle Rombaut*[†], *George Tziritas*[‡] *and William Lim*[†]

[†]Laboratoire des Images et des Signaux, 46 avenue Félix Viallet, 38031 Grenoble, FRANCE - @: *first_name.family_name*@lis.inpg.fr

[‡]Department of Computer Science, University of Crete, P.O. Box 2208, Heraklion, GREECE - @: {cpanag, tziritas}@csd.uoc.gr

ABSTRACT

Many applications are concerned by human action recognition notably in multimedia and more particularly for video retrieval and archival. Usual approaches focus on probabilistic methods and assume a still camera. In this paper, a method based on the Transferable Belief Model fusion process and considering a moving camera is proposed. In this framework, the affine camera motion estimation and temporal variations of three major human points are combined. The method is tested on videos of athletics meetings in which *running*, *jumping* and *falling* actions have to be recognized. Results show the validity of the method for action recognition.

1. INTRODUCTION

HUMAN MOTION ANALYSIS [1] is of key importance in many applications like video retrieval and archival, security surveillance, medical diagnosis, sports analysis and smart rooms. It involves human motion detection/estimation [2], coarse or fine body limbs tracking [3] and behaviour understanding [4]. This paper especially focuses on the last point.

Behaviour understanding consists in the *recognition* and *description* of human actions and activities: an *action* is defined as a significant change in the human behaviour whereas an *activity* describes an ordered sequence of actions. A recognition process is usually performed by comparing observations to examples. Many methods exist like Templates Matching [5], Hidden Markov Models [6] and Dynamic Bayesian Networks [7]. The methods provided by the literature are often based on the probability theory [8] benefiting from a well-founded and granted theoretical framework.

Inherently to a probability-based description, the recognition rate highly depends on the learning data set that has to be close enough to what has to be recognized. Furthermore, the time granularity has an important impact on the recognition and problems of time-shift and scale are difficult to overcome. These problems can be encountered when the disparity between humans performing actions and activities is important. For real applications, it is hard to imagine that actions and activities are performed with the same manner by several humans.

Features usually used by the previous methods reflect human body parts motion and are generally provided by tracking algorithms [3] either active [9] or passive [10]. The number of tracked points depends on the limbs involved in the actions, i.e. the level of detail. Intrinsic properties of these sensors are taken into account in probability-based methods by representing and handling data *imprecision*.

The *Transferable Belief Model* (TBM) [11] is another data representation which is well adapted when statistics are lacking and when description can be made by expert knowledge. This model manages imprecision but also explicitly expresses *doubt* and *conflict* between sources of data.

An original method is described for action recognition and based on the combination of tracked points and camera motion within the TBM framework. The estimation of the camera motion is used as a complementary information of tracking, assuming that the camera roughly follows the human whose actions are analyzed.

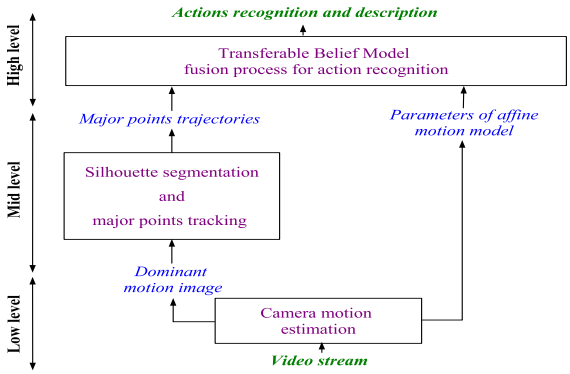

Figure 1: A bottom-up architecture to recognize action.

A general action recognition architecture is presented in section 2. Low and mid level modules, namely motion estimation and human points detection-tracking, are described in section 3. Section 4 is devoted to the higher level, namely action recognition. Videos of athletics meetings are used for testing: in this generic application, the purpose is to recognize *running*, *jumping* and *falling* actions in different jumps. Experimental results are shown in section 5 followed by the conclusion and future work.

2. ACTION RECOGNITION ARCHITECTURE

The synoptic of the system is presented in figure 1. It is based on two assumptions: first, the camera roughly follows the human and second, the evolution of head, trunk and end of legs positions give information on actions. Three levels of abstraction of the handled information are considered:

- *low level:* a camera motion estimation algorithm provides (i) six parameters, corresponding to the affine model of the camera motion, and (ii) an image whose pixel value depends on its belonging to the dominant motion. Figure 2(b) depicts an illustration.
- *mid level:* the dominant motion image represents the input of the segmentation module consisting of morphological operations and the resulting silhouette is used by a tracking algorithm to detect and trace three major points of the human body. A result is illustrated in figure 2(c).
- *high level:* affine motion parameters and tracking results are *combined* for increasing reliability and accuracy of action recognition. This combination is based on the Transferable Belief Model providing a *degree of belief* on each action and expressing conflict and doubt between data.

3. ACTION PARAMETERS

This section aims at describing the low and mid levels, i.e. modules providing temporal parameters used for recognition.

3.1 Camera motion estimation

It is assumed that the camera roughly follows the global motion of the human whose actions are analyzed. The algorithm presented in [12] is exploited to extract the camera motion. It consists in an iterative and robust multi-resolution estimation of parametric motion models between two successive frames. The affine motion model provides six parameters that are generally sufficient for most applications:

$$\begin{cases} v_x = a_0 + a_2 \cdot x + a_3 \cdot y \\ \\ v_y = a_1 + a_4 \cdot x + a_5 \cdot y \end{cases} \quad (1)$$

The parameters a_i represent the camera motion: horizontal and vertical translations (a_0 and a_1), rotation (a_4 and a_3) and divergence (a_5 and a_2). According to the previous assumption, these parameters correspond to the global motion of the human. Thus, if $a_0 > 0$ then the human horizontally translates towards the left. These parameters are then filtered to reduce noise (gaussian filter). A gray level image is also generated by the algorithm whose pixel value gives a piece of information on their belonging to the dominant motion, generally the motion of the background. In figure 2(b), the more a pixel is black, the less it belongs to the dominant motion and consequently, the more it belongs to the human.

3.2 Tracking

Three major points are tracked by an adaptation of [13]: *head*, *center of mass* and *end of leg*. They are sufficient to help in the recognition of global actions such as *running*, *jumping* and *falling*.

3.2.1 Silhouette segmentation

The tracking method needs a binary silhouette. The segmentation is based on the dominant image motion. A median filter and an opening are applied on the dominant image motion. The binary silhouette image is obtained by thresholding the previous result (the threshold is fixed empirically to 0.1). An illustration is given in figure 2(c) where the three major points appear.

3.2.2 Detection

The center of mass (x_c, y_c) of foreground pixels (F) is first computed. Then, the orientation Θ of the major human body axis passing through the mass center point is computed thanks to central moments $C_{1,1}, C_{2,0}, C_{0,2}$ defined as:

$$C_{p,q} = \sum_{(x,y)\in F} (x - x_c)^p (y - y_c)^q$$

$$\Theta = \arctan \frac{2C_{1,1}}{C_{2,0} - C_{0,2}}$$

It is assumed that the human stands in the first frame so that the head point (x_h, y_h) and the end of the leg (x_l, y_l) can be found. Both points represent extremities of the silhouette.

3.2.3 Tracking

The three major human points are then tracked. This procedure is executed in each frame of the sequence by considering the current frame and its previous. The current position of the points are first estimated as above and a reclassification of pixels of the binary silhouette image is performed in two steps. First, the minimum distance of each foreground pixel from the previous position of the three points is computed and if it is higher than a threshold (adaptive to image data and defined as a percentage of the human height), the foreground pixel is classified as the background. In the second step, background pixels belonging to human silhouette holes are classified as the foreground class. Afterwards, the three major human points are recalculated. An example is given in figure 2(c).

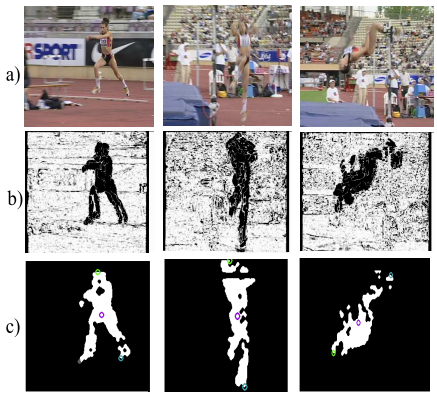

Figure 2: Results for a high jump sequence: (a) original frames for *running*, *jumping* and *falling* actions, (b) images of corresponding dominant motion and (c) silhouette segmentation the three major points.

4. FUSION AND RECOGNITION

The Transferable Belief Model (TBM) is a framework for the combination of sources of evidence. It was introduced by Smets and Kennes in [11] and comes from Dempster and Shafer's theory of evidence [14]. It is a mathematical model used to represent belief held by an agent about the value of the actual world. It seems well-adapted for action recognition because it explicitly expresses *doubt and conflict*: doubt expresses *gradualism* between actions and conflict reflects the need of improving the modeling and performing adaptations. The TBM is used as an alternative of probability-based methods for the fusion of the numerical parameters to find out actions in videos.

4.1 The Transferable Belief Model

The frame of discernment defines the possible hypothesis of the actual world. For instance, $\Omega_A = \{R_A, F_A\}$ is the frame of discernment of an action A, where R_A corresponds to *"A is Right"*, i.e. the current action is A, whereas F_A stands for *"A is False"* and means that the current action is not A. Hypotheses are exclusive and the frame of discernment is exhaustive.

For each frame, sensors give information concerning the real state of A. Each sensor S can be considered as a source of information about A. This information can be formalized as a Basic Belief Assignment (BBA) defined on 2^{Ω_A} by an application $m_S^{\Omega_A}$ from $X \in 2^{\Omega_A}$ to $m_S^{\Omega_A}(X) \in [0, 1]$. In the case concerned, X takes value in $2^{\Omega_A} = \{R_A, F_A, R_A \cup F_A\}$ where $R_A \cup F_A$ means *"A is Right* OR *False"*, i.e. the actual value of A is uncertain. The value $m_S^{\Omega_A}(X)$ represents the degree of evidential support that a specific element of Ω_A belongs to the set X, but not to a particular subset of X.

4.2 Definition of the BBAs

A BBA is defined for each sensor providing raw parameters. For that, *a priori* knowledge is used by expert observation of points' trajectories and camera parameters. These features are recalled in table 1.

The definition of the BBAs is drawn from *fuzzy rules*: when a raw parameter becomes available, inputs of the recognition module are updated by undergoing a numeric-to-symbolic conversion such as illustrated in figure 3. The

Camera motion (affine motion parameters)	
a_0	horizontal translation
a_1	vertical translation
a_2, a_5	divergence
a_3, a_4	rotation
tracking (coordinates)	
(x_c, y_c)	center of gravity
(x_h, y_h)	head
(x_l, y_l)	end of leg

Table 1: Raw parameters.

usual problem is to define the thresholds. They are currently set by expert knowledge but this point should be improved in future work.

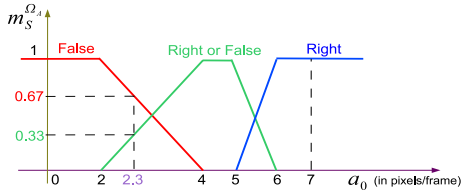

Figure 3: An example of a basic belief assignment based on fuzzy rules for the numeric-to-symbolic conversion of the horizontal motion parameter a_0.

For instance, consider figure 3 describing the conversion of the parameter of horizontal motion, noted a_0 in the sequel, for action $A =$ *running*. If $a_0 = 7$ (the unit is the number of pixels/frame), then all belief is given to R_A ($m_{a_0}^{\Omega_A}(R_A) = 1$) meaning that "$A =$ *running*" is true. As a second example, if $a_0 = 2.3$, $m_{a_0}^{\Omega_A}(R_A \cup F_A) = 0.33$ and $m_{a_0}^{\Omega_A}(F_A) = 0.67$. If $m_{a_0}^{\Omega_A}(R_A \cup F_A) > 0$, there exists a *doubt* between both *"Right"* and *"False"* proposals according to a_0.

4.3 Fusion

When several distinct sensors S_i give information about action A, the corresponding BBAs, $m_{S_i}^{\Omega_A}$, all defined on the same frame of discernment Ω_A, give a confidence on the reality of A. They can be combined by the conjunctive combination rule, noted $\textcircled{\cap}$, defined for two sensors such as:

$$(m_{S_1}^{\Omega_A} \textcircled{\cap} m_{S_2}^{\Omega_A})(X) = \sum_{\substack{X,Y,Z \subseteq \Omega_A \\ Y \cap Z = X}} m_{S_1}^{\Omega_A}(Y).m_{S_2}^{\Omega_A}(Z) \qquad (2)$$

When n distinct sensors S_i are available, the commutative and associative $\textcircled{\cap}$-rule provides a new BBA with $m_{S_{1,2\ldots n}}^{\Omega_A} = m_{S_1}^{\Omega_A} \textcircled{\cap} m_{S_2}^{\Omega_A} \ldots \textcircled{\cap} m_{S_n}^{\Omega_A}$. The $\textcircled{\cap}$-rule is used when several sensors have to agree about the reality of A. When sources are in conflict, a belief mass appears on the empty set, i.e. $m_{S_{1,2\ldots n}}^{\Omega_A}(\emptyset) > 0$. The conflict must be managed by analyzing inherent problems.

4.4 Action recognition

The recognition of *running*, *jumping* and *falling* actions is split into four steps. First, meta-parameters (table 2) are built. They represent relevant features for action recognition and they elucidate the description of actions. They are actually a function of the raw parameters. One of them is the human swing that describes how is positioned the main axis of the human compared to the horizontal axis of the frame. It is computed from the major points coordinates and brings information on the orientation of the human in the space (its value varies in $[0, 360]$ degrees). An illustration of an angle variation for a high jump sequence is given in figure 4.

Raw parameters	Meta-parameters
a_0, a_2	Horizontal motion
a_1	Vertical motion
(x_c, y_c), (x_h, y_h), (x_l, y_l)	Swing
(x_c, y_c), (x_h, y_h), (x_l, y_l)	Alternation of legs
y_c	Vertical variation

Table 2: Meta-parameters.

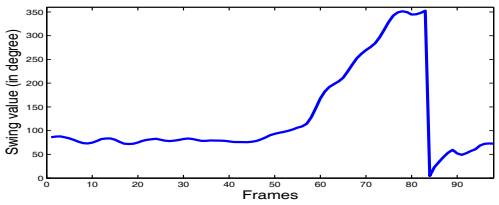

Figure 4: Evolution of the angle between human and horizontal axis computed with coordinates of the major points.

During the second step, BBAs are computed for all meta-parameters of table 2. The thresholds of the numeric-to-symbolic conversions are set once for each action and according to each jump. All frames of discernment correspond to $\Omega_A = \{R_A, F_A\}$, with A the action to recognize. In the third step, data fusion is performed from the BBAs expressed by each sensor. In the last step, temporal constraints on the reality of actions are added: if an action A is detected as *Right* after combination, i.e. $m_{S_{1,2\ldots n}}^{\Omega_A}(R_A) > 0$, then a potential change in the human behaviour appears. If its duration is at least Δ_A then the action is validated. Otherwise the belief is transferred into $R_A \cup F_A$, assuming that the range of frames concerned is uncertain. This process acts as a low-pass filter.

The actions are supposed to be completely independent of each other, i.e. the reality of one action gives no information about another one: one says that *actions are not exclusive* and therefore, they can happen at the same time. This phenomenon reflects a transition between actions, well-modeled within the TBM framework by the union of hypothesis. It can be noticed that new information concerning actions, e.g. *a priori* knowledge, can be easily taken into account into BBAs' computation by the TBM.

5. EXPERIMENTS

The proposed system is used to distinguish between *running*, *jumping* and *falling* actions in four different types of jump, namely high jump, long jump, pole vault and triple jump. The database consists in 33 videos and the number of frames concerning each action is given in table 3. Each video sequence is a 290×292 color avi video sequence of 25fps. The videos have been acquired under an unknown view angle and the most important movements of the camera are pan, tilt and zoom. It is assumed that only one people is moving however, the videos tested sometimes contain more people because athletes take part to a meeting. All the parameters are directly dependant of the camera motion estimation so, the quality of the videos is supposed to be sufficient enough to ensure its efficiency, notably concerning texture in image.

The setting of the numeric-to-symbolic conversions' thresholds are *set once for each action* and one setting is provided *for each type of jump*. Furthermore, the recognition is carried out frame by frame and *independently*.

The recognition results have been kept as belief masses because the real purpose of the proposed action recognition is to go on with activities so, the entire information has to be kept (a decision can annihilate precious details). However, it is necessary to know whether an action is true or not to assess the method. ***Decision*** – For that purpose, a

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	576	505	377	1458
Total	33	2410	1493	1038	4941

Table 3: Description of the database: *Running*, *Jumping* and *Falling* actions and their corresponding number of frames (col. 3-5). N_V is the total number of videos for each jump.

Jump/Action	Running		Jumping		Falling	
	$\mathcal{R}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{P}$
High jump	96.4%	79.6%	75.4%	73.3%	74.1%	91.6%
Long jump	90.0%	80.6%	58.2%	52.9%	64.8%	70.8%
Pole vault	81.9%	75.6%	74.5%	70.9%	72.8%	75.6%
Triple jump	85.9%	70.6%	55.6%	63.3%	62.9%	52.1%
Total	88.6%	76.7%	59.6%	66.1%	67.8%	64.0%

Table 4: Recall and precision of the recognition based on the credibility of *Running*, *Jumping* and *Falling* actions.

criteria based on the *credibility* of an action A, noted Cr_A, is tested. It considers A as true if $m^{\Omega_A}(R_A) > 0$. Other criteria can be chosen, for instance based on the plausibility or on the pignistic probability. However, Cr_A is chosen because it is a hard decision reflecting the quality of the recognition with a severe degree by focusing on the *specific* element R_A. ***Evaluation*** – 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 module by using the credibility-based criteria, and $C \cap R$ is the number of correctly retrieved frames.

Table 4 gathers the recall and precision indexes for each action and according to each type of jump. The last line represents their mean over all videos. ***Description*** – The running action is almost the same for each jump accounting for a high overall recognition rate. Jumping and falling recognition rates are less good than for running: confusing recall and precision, results are between 70.9% and 91.6% for the pole vault and the high jump, and between 52.9% and 70.8% for the long jump and the triple jump. ***Analysis*** – (i) Other moving people or objects used by the athlete like the perch for a pole vault disrupt the tracking and penalize the evaluation indexes: an automatic detection of the moving objects different to the athlete or a tracking coupled with a Kalman filter are possible solutions. (ii) An improvement of the recognition could be reached by a more detailed decomposition of these actions raising the problem of granularity. (iii) Actions have been described in a static way and the dynamic recognition is more relevant and challenging implying to take into account time and the chaining of events.

Figure 5 illustrates the evolution of the credibility of the three actions for a high jump sequence. The second *jumping* is due to the athlete expressing his happiness because he succeeded its attempt.

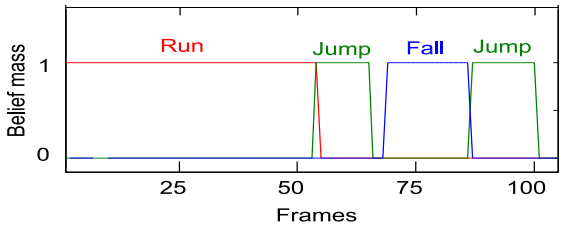

Figure 5: Actions recognition for a high jump sequence.

6. CONCLUSION

An original method based on the Transferable Belief Model was proposed. The data fusion took into account the camera motion and the trajectories of three major points of the human silhouette preliminarily segmented by image processing. These parameters were translated into basic belief assignments based on fuzzy rules. The database consisted in 33 athletics videos where the purpose was to recognize *running*, *jumping* and *falling* actions in four jumps namely high jump, long jump, pole vault and triple jump. An evaluation process based the credibility was applied and recall and precision indexes validated the method.

Work is under progress to integrate activity recognition into the proposed architecture, still based on the Transferable Belief Model. Adaptations by feedbacks between modules and the learning of the numeric-to-symbolic conversions' thresholds are foreseen.

ACKNOWLEDGMENTS

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

REFERENCES

[1] L. Wang, W. Hu, and T. Tan. Recent developments in human motion analysis. *Pattern Recognition*, 36(3):585–601, 2003.

[2] R. Cucchiara, A. Prati, and R. Vezzani. Real-time motion segmentation from moving cameras. *Real-Time Imaging*, 10:127–143, 2004.

[3] J. Wang and S. Singh. Video analysis of human dynamics – a survey. *Real-Time Imaging*, 9(5):321–346, 2003.

[4] J. K. Aggarwal and S. Park. Human motion: modeling and recognition of actions and interactions. In *3D Data Processing, Visualization, and Transmission*, 640–647, 2004.

[5] Y. Yacoob and M. Black. Parametrized modeling and recognition of activities. *CVIU*, 73(2):232–247, 1999.

[6] L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. of the IEEE*, 77(2):257–285, 1989.

[7] K. Murphy. *Dynamic Bayesian Networks: representation, inference and learning*. PhD thesis, UC Berkeley - Computer Science Division, 2002.

[8] S. Hongeng, R. Nevatia, and F. Bremond. Video-based event recognition and probabilistic recognition methods. *CVIU*, 96:129–162, 2004.

[9] V.J. Traver and F. Pla. Similarity motion estimation and active tracking through spatial-domain projections on log-polar images. *CVIU*, 97(2):209–241, 2005.

[10] G. Medioni, I. Cohen, F. Bremond, S. Hongeng, and R. Nevatia. Event detection and analysis from video streams. *PAMI*, 23(8):873–889, 2001.

[11] P. Smets and R. Kennes. The Transferable Belief Model. *Artificial Intelligence*, 66(2):191–234, 1994.

[12] J.M. Odobez and P. Bouthemy. Robust multiresolution estimation of parametric motion models. *J. of Vis. Comm. and Image R.*, 6(4):348–365, 1995.

[13] C. Panagiotakis and G. Tziritas. Recognition and tracking of the members of a moving human body. In *Articulated Motion and Deformable Objects*, 86–98, 2004.

[14] G. Shafer. *A mathematical theory of evidence*. Princeton University Press, Princeton, NJ, 1976.