Applied Soft Computing 46 (2016) 953–966

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

# Early detection of human actions—A hybrid approach

Ekta Vats, Chee Seng Chan *

*Centre of Image and Signal Processing, Faculty of Computer Science & Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia*

ARTICLE INFO

*Article history:*
Received 15 June 2015
Received in revised form 6 October 2015
Accepted 2 November 2015
Available online 21 November 2015

*Keywords:*
Human action recognition
Human activity recognition
BK subproduct
Human motion analysis

ABSTRACT

Early detection of human actions is essential in a wide spectrum of applications ranging from video surveillance to health-care. While human action recognition has been extensively studied, little attention is paid to the problem of detecting ongoing human action early, i.e. detecting an action as soon as it begins, but before it finishes. This study aims at training a detector to be capable of recognizing a human action when only partial action sample is seen. To do so, a hybrid technique is proposed in this work which combines the benefits of computer vision as well as fuzzy set theory based on the fuzzy Bandler and Kohout's sub-triangle product (BK subproduct). The novelty lies in the construction of a frame-by-frame membership function for each kind of possible movement. Detection is triggered when a pre-defined threshold is reached in a suitable way. Experimental results on a publicly available dataset demonstrate the benefits and effectiveness of the proposed method.



## 1. Introduction

Human action recognition has been widely studied over the years with real-time applications in video surveillance [1–3], health-care monitoring [4–6], sport analysis [7,8], etc. However, detecting ongoing human action early, i.e. as soon as it begins but before it finishes has not received much attention in the recent past. Most of the methods dealt with detection of the action after its completion. On the contrary, for early detection it is essential to detect partial action [9–13]. Early detection of human action is essential in several situations such as monitoring criminal activities, patients' fall detection, etc. Consider the example of an elderly care system in a hospital, it is crucial to accurately and rapidly detect the falling activity of the elderly patients as soon as possible, so that necessary medical care can be provided in a timely manner.

Early detection of human action is a daunting task given the vast amount of uncertainty involved therein. The conventional computer vision solutions often fall short of providing efficient solution as they are not robust enough to handle issues such as uncertainty, imprecision and vagueness. Hybrid techniques are believed to address these issues to a considerable extent by exploiting the strengths of one technique to alleviate the limitations of another [14,15]. Therefore, in this paper a hybrid technique for early detection of human action is proposed as the synergistic integration of computer vision solutions and fuzzy set theory. It is believed that computer vision methods and fuzzy approaches do not behave in a conflicting manner, but rather complimenting one another [16]. The fusion of these techniques towards performing human action recognition as early as possible can be achieved through proper hybridization. To this end, the relationship between a human and the action being performed is studied using the Bandler and Kohout's (BK) sub-triangle product (subproduct) [17], efficiently integrated with computer vision techniques including feature extraction and motion tracking to perform human action recognition effectively.

To the best of the authors' knowledge, this paper is the first attempt towards providing a solution to early human action detection using a hybrid technique, combining the benefits from computer vision and fuzzy set theory. Fuzzy BK subproduct is chosen in this work due to its flexibility and efficacy to be employed in real-world applications [18–21], and also its capability to imitate the natural human behavior, i.e. modus-ponen way [22]. Modus-ponen refers to our interpretation of available information while solving real-life problems, for example if $A$ implies $B$, and $A$ is asserted to be true, therefore $B$ must be true. Another issue addressed by the proposed method is to handle the cumulative tracking errors and precision problem using a set of overlapped fuzzy numbers known as fuzzy quantity space, where individual distance among them is defined by a predefined metric [23,24]. We intend to provide a solution for early human action detection closest to natural human perception. The novelty lies in the hybrid based learning formulation to train the early detector such that once the detector has been trained, it can be flexibly used in

* Corresponding author. Tel.: +60 3 7967 6433.
*E-mail addresses:* ektavats_2608@siswa.um.edu.my (E. Vats), cs.chan@um.edu.my (C.S. Chan).

http://dx.doi.org/10.1016/j.asoc.2015.11.007
1568-4946/

several ways depending upon the application. Experiments on a standard human action dataset illustrate the capability of the proposed hybrid technique to make reliable early detection of human action. The partial human action is modeled, where the fuzzy membership function provides the basis to detect an action before it is completed when a certain threshold is attained in a suitable way. To summarize, our main contribution is one of the first attempts to employ hybrid technique, a fuse between computer vision and fuzzy sets approaches for early detection of human action. Most of the classical solutions [16] had been focusing on human action recognition.

A preliminary version of this work was presented earlier [25]. The present work adds to the initial version in significant ways. Firstly, we improve the early human action detection framework by introducing fuzzy quantity space in the tracking stage to handle the cumulative tracking errors. Secondly, considerable new analyses and intuitive explanations are added to the original results. We also extend the original experiments from using a partial Weizmann dataset to a full Weizmann dataset.

This paper is structured in the following way. Section 2 provides the background on human action recognition from the fields of computer vision and fuzzy set theory. The BK subproduct approach is revisited in Section 2.2.1. Section 2.3 reviews early event detection. The proposed hybrid technique for early detection of human action is described in Section 3. Section 4 provides an analysis of the experimental results and assesses the effectiveness and benefits of the proposed method. Finally, Section 5 concludes the paper.

## 2. Background

This section gives an overview of the background of human action recognition, with reference to computer vision methods and fuzzy set oriented approaches with a short description of the BK subproduct inference mechanism. Fig. 1 represents a general framework for human action recognition, where for an input video, firstly, the human object is detected as low-level vision task, followed by human motion tracking in the mid-level processing. Furthermore, the literature on early detection of human action is reviewed with highlight on the state-of-the-art methods along with their limitations.

### 2.1. Human action recognition in computer vision

There exist several surveys of human action recognition in computer vision literature [26–29], focusing on various methods employed in the analysis of human body motion. Some of the recent works on human action classification includes [30–33], and for human activity recognition includes [34–37]. However, they are only capable of detecting complete human action. In the case of early human action detection, it is essential to detect partial action, as the concern is to recognize the activity being performed as soon as possible. Another limitation of these works is their inability to handle the uncertainties that exist in a real-world environment, which is taken into account by fuzzy approaches.

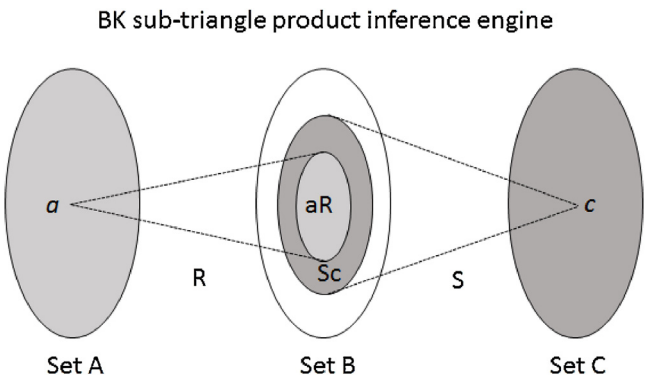



**Fig. 2.** Overview of BK subproduct: element $a$ in set $A$ is in relation with element $c$ in set $C$ if its image under $R$ ($aR$) is a subset of image $Sc$.

### 2.2. Fuzzy human action recognition

In recent times, the fuzzy approaches such as type-1 fuzzy inference system [38,39], fuzzy HMM [40] and hybrid techniques [14,15],have proven to be beneficial in human action recognition. Fuzzy human action recognition techniques can be efficiently used to distinguish the human motion patterns, and recognize the human activities with their capability to model the uncertainties involved therein. Nonetheless, fuzzy vector quantization [41] and qualitative normalized template [24,42] provide the capability to handle the complex human activities occurring in everyday life. However, these approaches are tailor-made for human action recognition and classification tasks only, and lacking in ability to detect an action early.

BK relational products have been successfully employed in developing the inference engine for several applications such as in the medical expert system [43], information retrieval [44], autonomous underwater vehicles' path navigation [19], land evaluation [20], scene classification [21,45], etc. In this paper, a hybrid technique of fuzzy BK subproduct and the computer vision solutions is employed for human action recognition. In order to provide a better understanding of the concept, the following section revisits BK subproduct.

#### 2.2.1. BK subproduct revisit

Bandler and Kohout [17] proposed that the relationship between two indirectly associated sets can be studied with the BK relational product that defines the relationship between the elements within the two indirectly associated sets as the overlapping of their images in a common set. Fig. 2 gives an overview of the BK subproduct for crisp relations. Let us assume that there exist three sets: set $A = \{a_i | i = 1, \cdots, I\}$, set $B = \{b_j | j = 1, \cdots, J\}$ and set $C = \{c_k | k = 1, \cdots, K\}$. If a relation $R$ is defined between $A$ and $B$ such that $R \subseteq \{(a, b) | (a, b) \in A \times B\}$, and a relation $S$ is defined between $B$ and $C$ such that $S \subseteq \{(b, c) | (b, c) \in B \times C\}$, then the BK subproduct can be defined as:

$$R \lhd S = \{(a, c) | (a, c) \in A \times C \quad \text{and} \quad aR \subseteq Sc\} \tag{1}$$

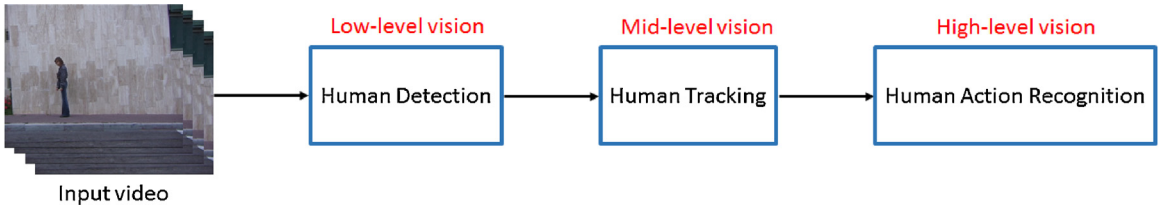



**Fig. 1.** A general framework for human action recognition.

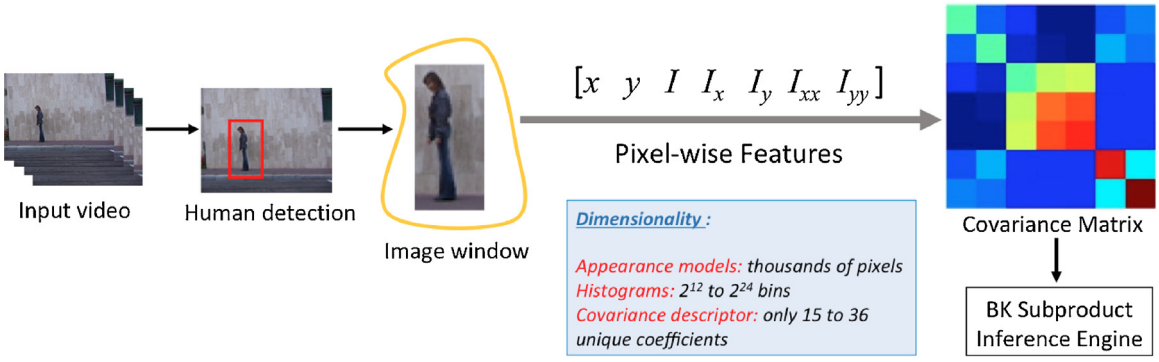



**Fig. 3.** An example of human motion image.

BK subproduct finds all $(a, c)$ couples such that the image of $a$ under relation $R$ in $B$ ($aR$) is among the subset of $c$ under the converse relation of $S$ in $B$ ($S^c$), as illustrated in Fig. 2. As an extension to crisp BK subproduct, [17] proposed fuzzy BK subproduct to handle the uncertainty issues that exist in the real-world. Observing Eq. (1), it can be seen that $aR \subseteq Sc$ is the main element to retrieve the relationship between $a$ and $c$. Therefore, the fuzzy subsethood measure was developed in [46] based on the fuzzy implication operators '$\rightarrow$'.

Let $P$ and $Q$ be the fuzzy subsets in the universe $X$, such that $x \in X$. Then the possibility that $P$ is a subset of $Q$ is given as:

$$\pi(P \subseteq Q) = \bigwedge_{x \in X} (\mu_P(x) \rightarrow \mu_Q(x)) \tag{2}$$

where $\bigwedge$ represents the arithmetic mean in mean criterion or the infimum operator in harsh criterion; $\mu_P(x)$, and $\mu_Q(x)$ represents the membership function of $x$ in $P$ and $Q$ respectively; while $\rightarrow$ is the fuzzy implication operator. Utilizing Eq. (1) and (2), BK subproduct as the composition of relations between $a_i \in A$ and $c_k \in C$ is defined as follows [17]:

$$R \lhd S(a, c) = \bigwedge_{b \in B} (R(a, b) \rightarrow S(b, c)) \tag{3}$$

where, $R(a, b)$ represents the membership function of the relation $R$ between $a$ and $b$; and $S(b, c)$ represents the membership function of the relation $S$ between $b$ and $c$.

BK subproduct is a flexible approach that can be applied in real-life applications. Consider an example of a human motion image, as illustrated in Fig. 3 where an actor performs an action. Given an input video of action sequences, the human object is first detected for each image frame. This is followed by feature extraction. Features are the elements to be modeled and represented in a meaningful manner to signify the action. A popular feature extraction approach is to represent the image window by a covariance matrix of features [47], where the concept of covariance implies how much two variables vary together.

Let $f_i$ denote the features extracted from image frames $i = 1, \cdots, I$ of the video describing the human action $a_k$, for $k = 1, \cdots, K$ action classes. The features extracted can be associated directly with the pixel coordinates. Therefore, the pixel-wise features $f_i = [x\, y\, I\, I_x\, I_y\, I_{xx}\, I_{yy}]$ can be extracted as represented in Fig. 3. By constructing the covariance of different features of a human image window (e.g. color, gradient, motion, edge etc.), the information from the histograms and the appearance models can be extracted. And by using bag of covariance matrices, the detection of actions, poses and shape changes can be taken into account efficiently [47].

To detect human action in a given image, a BK subproduct classifier is first trained. The indirect relationship between the features representing the human image and the actions being performed can be deduced using fuzzy BK subproduct. This is conditional on the presence of an intermediate set that is in relation with both $f_i$ and $a_k$, such as the human body part-based model $m_j$ for $j = 1, \cdots, J$ (where $J$ denotes the number of models), obtained as a result of covariance tracking. The BK subproduct classifier is invoked at each candidate image window to determine the target human action. The detection is triggered at frame $i$ when the detector obtains the segment having the highest membership value.

For testing image sequences, this entails finding the features that signify the desired human action. There exists several popular methods to perform this task e.g. using the well-known classifiers such as SVM, KNN and so on, but the employability of these algorithms depends on the desired application and its requirements. Although BK subproduct is not a popular classifier, but it can be efficiently used for classification tasks with the ability to provide a solution closer to how human interpret a situation in real life. For example, the relationship between a set of features and the action classes can be established if there exists an intermediate element that is in relationship with both, such as a human body model generated as a result of human motion tracking.

Furthermore, the introduction of fuzzy subsethood measure in the BK subproduct simplifies the classification process in the sense that the crisp BK subproduct allows an action to belong to a single class only i.e. mutually exclusive classification approach. Whereas, fuzzy BK subproduct provides flexibility where an action can belong to a particular class with a certain degree of belongingness defined using fuzzy membership functions, and therefore offers non-mutually exclusive action classification. This is very crucial for early detection of human actions, because initially there is no information available about the action being performed. As the video progresses, the membership function values generated using BK subproduct vary following a certain trend (e.g. monotonically increasing or decreasing), enabling frame-by-frame action classification.

### 2.3. Early event detection

In general, there exist several human action recognition methods in literature [16,48]. Most of the methods dealt with detection of action after its completion. For early detection it is essential to detect partial human action [9,49,10,50,11–13,51]. The following subsection discusses the pros and cons of the existing methods for early human action detection, followed by a short review on the learning mechanism for early event detectors in Section 2.3.2.

#### 2.3.1. Pros and cons of the existing methods

Most of the existing work dealing with early detection of human actions aims at detecting unfinished activity. [9] proposed the integral bag-of-words and dynamic bag-of-words approaches as an extension to the bag-of-words paradigm for early recognition of ongoing human activities, and delivered promising results. However, the model learned for activity recognition may not be representative if the action sequences of the same action class

have large appearance variations. Also, it was found to be sensitive to outliers. The solution to these two issues was provided in [49] where the action models were built by utilizing sparse coding to learn the feature bases, and using the reconstruction error in the likelihood computation. Other limitations of [9] include the assumption made that the activities within the same action class always have identical speed and duration which is not true in most cases. Also, the poor discriminative model generated to describe human action, ignoring the bag-of-words model in spatial-temporal relationships among the interest points. This issue was taken into account in [10] where a spatial-temporal implicit shape model was proposed to model the relationship between the local features, and at the same time predict multiple activities. The method proposed in [50] incorporated an important prior knowledge that as new observations are available when the action video progresses, the amount of crucial information about the action also increases. However, the methods [9,49,10] did not utilize this prior knowledge. In addition, [50] modeled the label consistency of segments, which provides discriminative local information, as well as implicitly captures the context-level information that is useful for predicting actions. Moreover, [50] captured the action dynamics in both global as well as local temporal scales, unlike [9,49] where the dynamics in single scale were captured. In spite of the advantages these methods offer, they lack in the ability to handle the uncertainties that exist in a real-world.

The early recognition of human action for the dynamic first-person videos was studied in [11], where the pre-activity observations were considered that includes the frames 'before' the starting time of the activity. However, this work is different from the goal of this paper. ARMA-HMM based approach was employed in [12] which integrates both the predictive power of sequential model HMM (Hidden Markov model) and the time series model ARMA (Autoregressive-moving-average). Unfortunately, it requires building separate HMMs for each activity and therefore is computationally expensive. Max-Margin Early Event Detector (MMED) was proposed for early detection of events in [13,51] which is based on the Structured Output SVM [52] and requires extensive labeling on each of the training samples. In terms of timeliness and accuracy, MMED performs efficiently. However, early detection of human action is a complex task given the vast amount of uncertainty involved therein. An efficient algorithm should be able to handle even the minutest level of uncertainty for a reliable decision making.

The conventional computer vision solutions [9,49,10,50,11–13,51] often fall short of providing an effective solution as they are not robust enough to handle issues such as uncertainty, imprecision and vagueness that arise in a real-world. The fuzzy approaches are well-known in offering an effective solution with the inherent capability of assigning a degree of belongingness to a human action using the fuzzy membership function. The problem of early human action recognition can be efficiently addressed by integrating computer vision solutions with fuzzy set oriented techniques in a way that the strength of fuzzy set theory can alleviate the limitation of computer vision solutions. In the following section, the learning mechanism for early event detectors is reviewed, which forms the baseline for designing the learning formulation for early human action detector.

#### 2.3.2. Review on learning mechanism for early event detectors

For early event detection, partial events are used as positive training examples [13], instead of a complete event. For a training set $X^i$ of length $l^i$ and time $t = 1, 2, \cdots, l^i$, the output of the detector at time $t$ is a partial event represented as:

$$g(X^i_{[1,t]}) = y^i_t = \underset{y \in Y(t)}{arg\,max} f(X^i_y) \tag{4}$$

where, $y^i_t = y^i \cap [1, t]$ is the part of event $y^i$ that has already happened and is possibly empty; $g(X^i_{[1,t]})$ is the output of detector on the subsequence of time series $X^i$, not the entire set; and $f(X^i_y)$ is the detection score function. It is required that the detector score function is a monotonic and non-decreasing function. This means that the score of the partial event $y^i_t$ should be greater than the score of any segment $y$ ending before the partial event, which has been seen in the past, i.e.

$$f(X^i_{y^i_t}) \geqslant f(X^i_y) + \Delta(y^i_t, y) \forall y \in Y(t) \tag{5}$$

where $\Delta(y^i_t, y)$ is the loss of detector for outputting $y$ when the desired output was $y^i_t$.

The constraint in Eq. (5) is enforced for all $t = 1, 2, \cdots, l^i$. The learning formulation for early event detector is obtained as [13]:

$$\min_{w,b,\xi^i \geqslant 0} \frac{1}{2}\|w\|^2 + \frac{C}{n}\sum_{i=1}^{n} \xi^i \tag{6}$$

so that

$$f(X^i_{y^i_t}) \geqslant f(X^i_y) + \Delta(y^i_t, y) - \frac{\xi^i}{\mu\left(\frac{|y^i_t|}{|y^i|}\right)} \quad \forall i, \forall t = 1, \ldots, l^i, \forall y \in Y(t) \tag{7}$$

where $w$ is a weight vector, $b$ is a scalar bias term, $C$ is the cost parameter, and $n$ denotes the number of instances of the training data. This is an extension of Structure Output SVM, with the alteration on setting $t = 1, 2, \cdots, l^i$ instead of $t = l^i$, because partial events are trained instead of a complete event. An additional slack variable $\xi^i$ is added as a rescaling factor for correctly detecting the occurrence of an event at time $t$.

## 3. Hybrid technique for early detection of human action

Human action recognition is a high-level computer vision problem, which involves human detection in the low-level processing and human motion tracking in the intermediate level. However, vast amount of uncertainty, imprecision and vagueness issues may exist that are required to be dealt with using an efficient algorithm (e.g. fuzzy approaches). Therefore, a hybrid technique for early detection of human action is proposed in this work as the integration of computer vision solutions as well as fuzzy set theory, where the hybridization is performed on the tracking output generated and the fuzzy BK subproduct. Fig. 4 highlights the overall pipeline of the proposed hybrid solution and will be discussed step-by-step in this section.

### 3.1. Feature extraction

Given an input video action sequences, the object window is represented as a covariance matrix of features following the method in [47]. The covariance matrix is a symmetric matrix where the diagonal represents the variance of each feature in the image, and their representative correlations is represented by the non-diagonal. The covariance matrix is scale-invariant and efficiently combines multiple features without the need to normalize features or blend weights. The reason for choosing this method is to capture the spatial as well as statistical properties along with their correlation within the same representation. Let $F$ be the $W \times H \times d$ dimensional RGB feature image of an image $I$, such that $F(x, y) = \Phi(I, x, y)$ where the function $\Phi$ can be any mapping such as image gradients, color, edge magnitude or orientation. Let $f_{ii=1..I'}$ be the d-dimensional feature vector inside a rectangular window $R'$ where $R' \subset F$. A feature

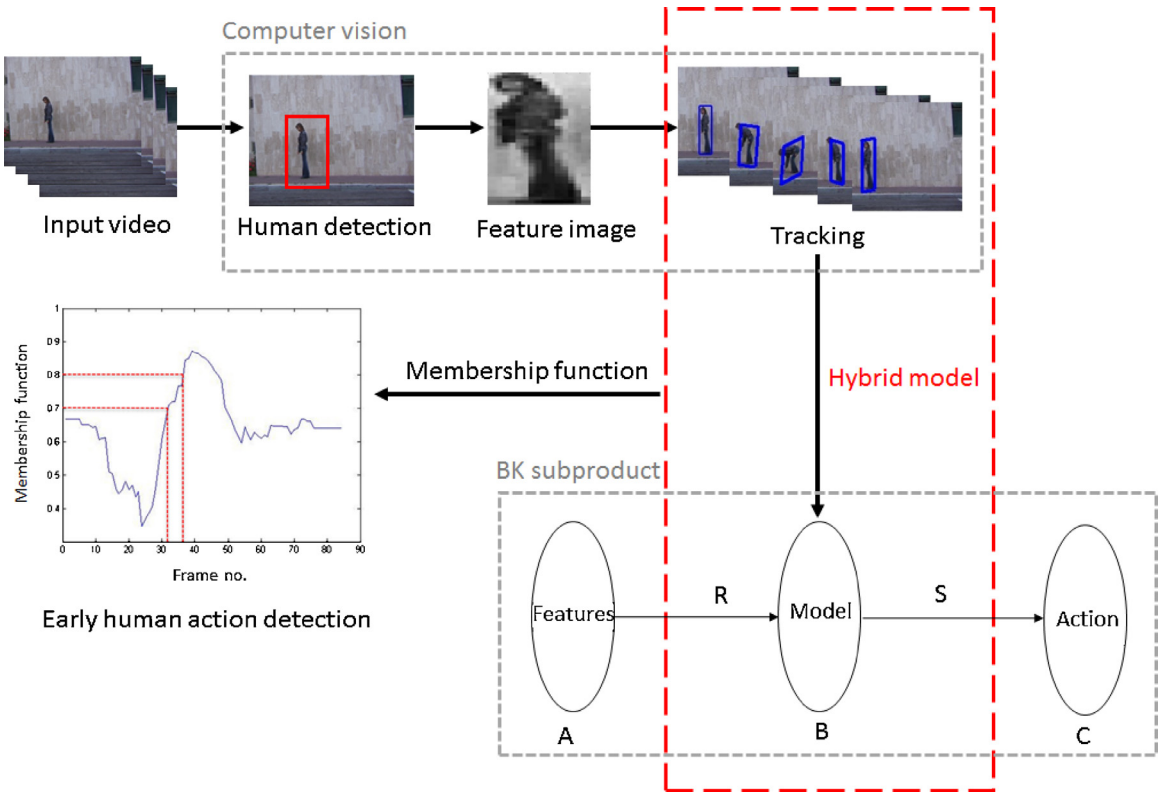



**Fig. 4.** Overall pipeline of the proposed hybrid technique. The hybridization is performed on the tracking output from computer vision solutions and the set B of fuzzy BK subproduct which includes a set of human body part-based models obtained from the human motion tracking. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

vector $f_i$ is constructed using: (i) spatial attributes based mapping - obtained from the pixel coordinates values, and (ii) appearance attributes based mapping - e.g. gradient, color or infrared. The features extracted may be associated directly with the pixel coordinates ($f_i = [x\, y\, I(x, y)\, I_x(x, y) \ldots]$), or can be arranged in a radially symmetric relationship ($f_i^r = [\|(x', y')\|\, I(x, y)\, I_x(x, y) \ldots]$).

### 3.2. Covariance tracking

Tracking is important in finding the correspondences between the previously detected objects in the current image frame. A common approach in tracking is to employ predictive filtering where the object's location in the distance calculation and color attributes are used to update the model [53]. When the measurement noise is assumed to be Gaussian, the Kalman filter offers an optimal solution. Whereas Markovian filters can be applied for tracking when the state space consists of a finite number of states. Another well-known approach is to employ particle filters which are based on Monte Carlo integration methods. In particle filtering, the current density of the state (i.e. speed, size, location) is represented using a set of random samples with associated weights. Further, the new density is computed utilizing these samples and weights. However, the main disadvantage of particle filtering is that it is based on random sampling, and therefore suffer from the problem of sample degeneracy and impoverishment, especially for the higher dimensional representations [47].

In order to find a global optimal solution, the covariance tracking method [47] is employed. It is a simple algorithm used to track non-rigid objects using covariance based object description. A model update mechanism is incorporated using Lie algebra to adapt to the undergoing object deformations and appearance changes. Unlike other tracking methods, covariance based tracking does not make any assumption on the measurement noise as well as the motion of the objects tracked, and has shown remarkable detection accuracy for the moving objects in non-stationary camera sequences. Covariance tracking is performed as follows.

For a given object region $R'$, a $d \times d$ covariance matrix of features $C_{R'}$ is computed as the model of the human object:

$$C_{R'} = \frac{1}{MN} \sum_{i=1}^{MN} (f_i - \mu_{R'})(f_i - \mu_{R'})^T \tag{8}$$

where, $\mu_{R'}$ is the vector of the mean of the corresponding features for the points in region $R'$. A single covariance matrix extracted from a region is sufficient to perform matching of the region in multiple views and poses. In the current image frame, the region having the minimum covariance distance from the model is located and assigned as the estimated location.

Furthermore, the covariance tracking algorithm is modified to perform part-based human motion tracking. Human body is segmented into three parts: head, torso and leg, and the covariance tracking is performed on each of the part, resulting in a part-based model $m$. Five distinct models were generated: (i) head distance – model the head movement from start to end frame, (ii) body distance – model the position changes of the human body from the first frame, (iii) leg distance – model the distance between both legs, (iv) hand distance – model the hand movement from start to end frame, and (v) ground distance – model the distance of the human body from the ground. In order to adapt to variations, a set of previous covariance matrices are kept and an intrinsic mean is extracted using Lie algebra [47].

It is crucial for the tracking algorithm to not suffer from problems such as tracking precision issue due to the position changes of each body part (head, torso and leg) evolving over time, or the cumulative errors generated because of the uncertainties arising due to different height, size and step size of each human. These

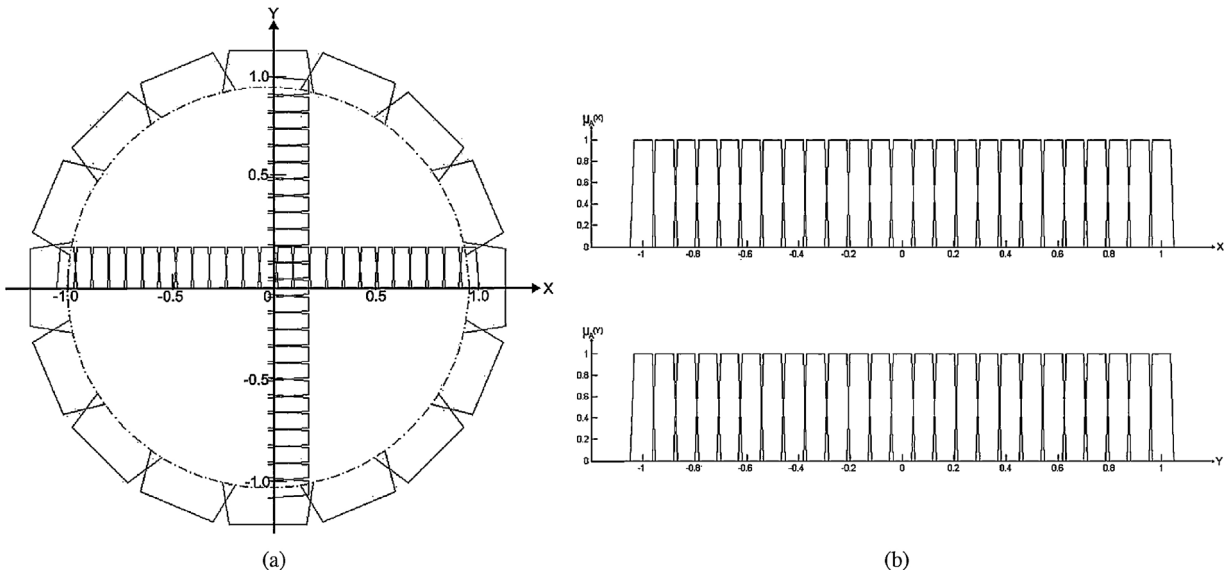



**Fig. 5.** (a) Description of the Cartesian translation in the conventional unit circle replaced by the fuzzy quantity space. (b) Element of the fuzzy quantity space for translation $(X, Y)$ in the fuzzy qualitative unit circle as a finite and convex discretization of the real number line [24].

problems can directly affect the performance of the higher level task. Therefore, the tracking output is fuzzified using fuzzy qualitative quantity space as discussed in the following section.

### 3.2.1. Fuzzy qualitative quantity space

The fuzzy qualitative quantity space can be defined as a set of overlapped fuzzy numbers whose individual distance among them is defined by a predefined metric [23,24]. Four tuple fuzzy numbers $[a, b, \alpha, \beta]$ are employed to describe each state in the fuzzy qualitative unit circle (Fig. 5(a)) that is a finite and convex discretization of the real number line. In this paper, the main motivation behind employing the fuzzy qualitative unit circle is to model the accumulated errors due to the position changes of each body part (head, torso and leg) evolving over time. Besides that, this approach can help in dealing with the tracking errors and precision problem because of the uncertainties arise due to different height, size and step size of each human.

In the proposed method, the rigid motion of each body part is represented using the fuzzy qualitative translation states. A fuzzy qualitative unit circle as shown in Fig. 5 is constructed using Eq. (9), following the approach in [23,24]:

$$\lim_{s \to s_0 = 10} C_t(s) = QS(qp_l) \tag{9}$$

where the translation component in the conventional unit circle is replaced by the fuzzy qualitative quantity space and $s$ denotes the number of states representing the $x - y$ translation employed in the quantity space to represent the fuzzy qualitative unit circle. Empirically, the translation was selected as $s = 10$. The fuzzy qualitative quantity space $Q$ consists of the translation component $Q^d$ represented as:

$$Q^d = QS_d(l_j),\ for j = 1, 2, \ldots, n \tag{10}$$

where $QS_d(l_j)$ denotes the state of a distance $l_j$, and $n$ represents the number of elements in the translation component. The final output generated is the fuzzified tracking result, normalized using the fuzzy qualitative quantity space with values between 0 and 1.

## 3.3. Hybrid model

The output from the human body part-based covariance tracking, normalized using fuzzy qualitative quantity space is integrated with the fuzzy BK subproduct with proper hybridization process to perform human action recognition as presented in Fig. 4.

Given an input action video, let $A = \{f_i | i = 1, \cdots, I\}$ denote the set of features extracted from image frames $i$ of the video describing the human action. Let set $C = \{a_k | k = 1, \cdots, K\}$ be the set of human action. $A$ has no direct relation with $C$, since there is no information about which action is being performed and by whom. However, if there exists an intermediate set $B$, which is in relation with both $A$ and $C$, the indirect relationship between $A$ and $C$ can be derived using fuzzy BK subproduct, and utilize this information to detect an action as early as possible. Therefore, let set $B = \{m_j | j = 1, \cdots, J\}$ constitute the human body part-based model, obtained as a result of covariance tracking. Using this intermediate set, the relationship between image features $f$ in set $A$ and the action $a$ in set $C$ can be therefore obtained by rewriting Eq. (1) as:

$$R \lhd S = \{(f, a) | (f, a) \in A \times C \quad \text{and} \quad fR \subseteq Sa\} \tag{11}$$

where $fR \subseteq Sa$ is the main element in retrieving the relationship between $f$ and $a$, and is obtained from the covariance tracking. The composition of relation between $f_i \in A$ and $a_k \in C$ can be defined using the fuzzy subsethood measure as follows:

$$BK: \quad R \lhd_{BK} S(f, a) = \frac{1}{J} \sum_{m \in B} (R(f, m) \rightarrow S(m, a)) \tag{12}$$

where, $R(f, m)$ represents the membership function of the relation $R$ between $f$ and $m$, and $S(b, c)$ represents the membership function of the relation $S$ between $m$ and $a$. Therefore, Eq. (12) represents the hybrid model mathematically. The hybrid model performs the integration of the models obtained from human motion tracking into the intermediate set $B$ of BK subproduct. As a result, set $B$ includes five distinct models $m_1 - m_5$ generated from the covariance tracking, i.e. head distance, body distance, leg distance, hand distance, and ground distance.

For each image frame, the membership function values generated from Eq. (12) are modeled for early detection of human action. For example, for an action video with $n$ number of frames, invoking BK subproduct inference engine for each frame will yield a membership function value for each frame as an output. The early detector models the frame-by-frame membership function values generated from BK subproduct and triggers an action when it exceeds a pre-defined threshold monotonically. Even if a single

action is being continued, the membership grades are constructed using BK subproduct (Eq. (12)) frame-by-frame, and the early detector detects the action in a similar manner. When the membership function value attains the desired threshold value at a certain frame, the detector stops, and triggers the action at that particular frame number. Section 3.4 explains the overall process in detail.

### 3.4. Early detection of human action

Early detection of human action involves processing in real-time. The detector reads from a stream of input video and keeps a sequence of observations in the memory, continuously monitoring the occurrence of the target action. If the target action is detected, the frame number at which the detector triggers is returned.

However, in order to detect an action as early as possible, partial actions are used as positive training examples, instead of a complete action sequence. Let $(X^1, y^1),\cdots,(X^n, y^n)$ be the set of a series of actions performed by a human and the associated ground truth annotations for the action of interest such that $y^i = [s^i, e^i]$, where $s^i$ denotes the start of the action and $e^i$ denotes the end of the action in the time series of action $X^i$. Let $t_0$ denote the beginning of the action video, and the length of the partial and complete action that the detector needs to detect be bounded by $l_{min}$ and $l_{max}$. Let $Y(t_0, t)$ denote the set of length-bounded time intervals from time $t_0$ to time $t$. Also, for a time series of action $X$ of length $l$, let $Y(l)$ denote the set of all possible locations of an action in a video. For an interval $y = [s, e] \in Y(l)$, let $X_y$ denote the subsegment of $X$ from frame $s$ to $e$ inclusive. Then, the output of detector that is the segment having the highest membership value (degree of belongingness to an action) is represented as:

$$D(X^i_{[t_0,t]}) = y^i_t = \underset{y \in Y(t_0,t)}{arg\,\max} \mu(X^i_y) \tag{13}$$

where, $D(X^i_{[1,t]})$ denotes the output of detector on the subsequence of $X^i$ from the initial frame to the $t^{th}$ frame only, instead of entire $X^i$. If $D(X^i_{[t_0,t]}) = \{\emptyset\}$, no action is detected. $\mu(X^i_y)$ represents the membership function of the segment $X^i_y$ belonging to the time series of action $X^i$. Similarly, the detector's output at $t+1$ can be computed as:

$$D(X^i_{[t_0,t+1]}) = y^i_{t+1} = \underset{y \in Y(t_0,t+1), y(2)=t+1}{arg\,\max} \mu(X^i_y) \tag{14}$$

where $y^i_{t+1}$ is the segment that attains the maximum membership function at $t+1$. The overall computational cost involved for the detection is $O(l)$, where $l_{min} \le l \le l_{max}$.

For early human action detection, it is desirable for the membership function $\mu(X^i_y)$ to be monotonic and non-decreasing as presented in Fig. 6, using the example of *bend* action. This means that the membership function of the partial action $y^i_t$ should always be higher than the membership function of any segment that ends before the partial action [13]. Therefore, Eq. (13) must hold with the desired property:

$$\mu(X^i_{y^i_t}) \geqslant \mu(X^i_y) \forall i, \forall t = 1, \ldots, l^i, \forall y \in Y(t) \tag{15}$$

The constraint shown in Eq. (15) is enforced for all $t = 1, 2, \cdots, l^i$, instead of $t = l^i$ as the partial actions are being trained instead of a complete action. The learning formulation for early human action detection is obtained as in Eq. (13)-(15), where the membership function $\mu(X^i_y)$ is learned using the proposed hybrid technique. In this paper, the target action of multiple classes is detected. Therefore, the detectors are trained and used separately for each of the target action classes. The challenge is to study the indirect relationship between the actor and the action being performed in the video, modeling the frame-by-frame arrival of data, and subsequently perform action classification on the basis of the membership function values generated from the hybrid model. Therefore, Eq. (12) can be re-written as:

$$\mu(X^i_y) = R \triangleleft_{BK} S(f, a) = \frac{1}{J} \sum_{m \in B} (R(f, m) \rightarrow S(m, a)) \quad \forall i, \forall t = 1, \ldots, l^i, \forall y \in Y(t) \tag{16}$$

where Eq. (16) yields the desired membership function required for early detection of human action. When the membership function value monotonically exceeds a pre-defined threshold, the detector triggers the action.

Early detection of human action can also be defined in terms of the semantic relationship between human and the action. Given an input set of training time series of action sequences $X^1$, $X^2$,..., $X^n$ performed by a human and the associated ground truth annotations $y^1$, $y^2$, $\cdots$, $y^n$ for the action of interest, it is assumed that each training action sequence contains at most one action of interest, as a training sequence containing several actions can always be divided into smaller subsequences of a single action. Therefore, $y^i = [s^i, e^i]$ consists of two numbers that indicate the start and end of the action in the time series of action $X^i$ respectively. Early detection of human action aims at finding the semantics (human – action) in a set of series of actions $(X^1, y^1), \cdots, (X^n, y^n)$ where $y^i \subset [s^i, e^i]$. However, the semantics (human – action) remain invariant if all the frames have been used. If so, a Silico DNA based computing is considered to serve the purpose effortlessly. For example, in [54] a DNA based computing approach for understanding complex shapes have been proposed where the authors have shown that whatever may be the outlook of the image frames, they underlie the same semantics (fern-leaf).

However, the method in [54] is applicable to two-dimensional image data only. Our proposed method is well-equipped to handle these issues, in the sense that there cannot possibly exist a situation where all the frames have been used to detect an action as then it will be same as the conventional classification problem which requires seeing a complete action. Instead, our early detector is trained to detect partial actions. This means that for an interval $y = [s, e] \in Y(l)$, where $Y(l)$ denote the set of all possible locations of an action in a video, and $X_y$ denote the subsegment of $X$ from frame $s$ to $e$ inclusive, the detector $D(X^i_{[t_0,t]})$ outputs the segment having the highest membership degree of belongingness to an action i.e. $\mu(X^i_y)$, which is a partial segment $y^i_t$ instead of a complete action $y$. Furthermore, Eq. (15) can be modified by adding an additional variable $\Delta(y^i_t, y)$ which is the loss of detector for outputting $y$ when the desired output is $y^i_t$, represented as follows:

$$\mu(X^i_{y^i_t}) \geqslant \mu(X^i_y) + \Delta(y^i_t, y), \forall i, \forall t = 1, \ldots, l^i, \forall y \in Y(t) \tag{17}$$

where $\Delta(y^i_t, y)$ handles the exception where all the frames have been used and the detector fails to detect the occurrence of an action early.

#### 3.4.1. Impact of fuzzy implication operators

An important property of Eq. (16) to be taken into consideration is which implication operator '$\rightarrow$' to use to infer the relation '$R(f, m) \rightarrow S(m, a)$'. There exists a number of fuzzy implication operators in the literature (Table 1), but are tailor-made for specific applications. A third dimension 'time' plays a crucial role in human action recognition to determine how human movement changes over time. Therefore, a small modification on the popular implication operators i.e. Łukasiewicz ($p \rightarrow_{Ł} q$) and Kleene-Dienes ($p \rightarrow_{KD} q$) operators was done to accommodate 'time' as an additional dimension as follows:

$$p \rightarrow_{\text{new}Ł} q = \min(1, 1 - p_t + q_t), \forall i, \forall t = 1, \ldots, l^i \tag{18}$$

$$p \rightarrow_{\text{new}KD} q = \max(q_t, 1 - p_t), \forall i, \forall t = 1, \ldots, l^i \tag{19}$$

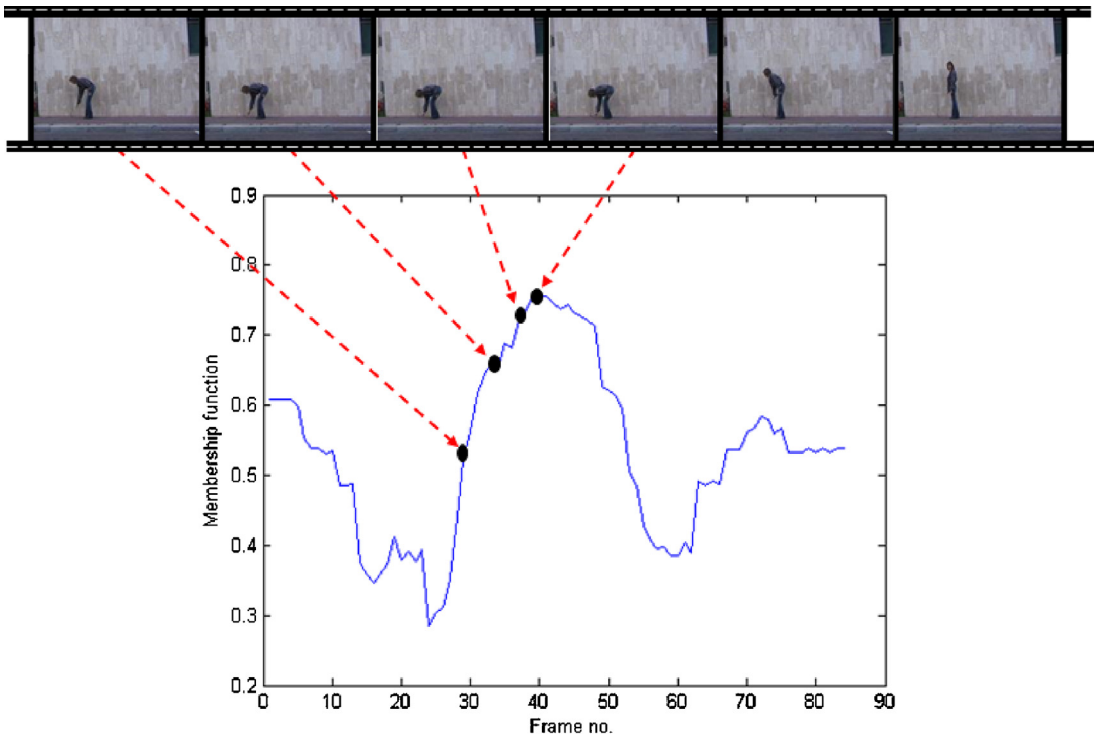



**Fig. 6.** Monotonicity requirement: the membership function of the partial action should always be higher than the membership function of any segment that ends before the partial action.

**Table 1**
Some fuzzy implication operators with their symbolic representation and definitions [18].

| Implication operator | Symbol | Definition |
|---|---|---|
| S# - Standard Sharp | $r \rightarrow_{S\#} s$ | $\begin{cases} 1 & \text{iff } r \neq 1 \text{ or } s = 1 \\ 0 & \text{otherwise} \end{cases}$ |
| S - Standard Strict | $r \rightarrow_{S} s$ | $\begin{cases} 1 & \text{iff } r \leq 1 \\ 0 & \text{otherwise} \end{cases}$ |
| G43 - Gaines 43 | $r \rightarrow_{G43} s$ | $\min(1, \frac{r}{s})$ |
| KD - Kleene-Dienes | $r \rightarrow_{KD} s$ | $\max(s, 1-r)$ |
| R - Reichenbach | $r \rightarrow_{R} s$ | $1 - r + rs = \min(1, 1 - r + s)$ |
| L - Łukasiewicz | $r \rightarrow_{Ł} s$ | $\min(1, 1 - r + s)$ |
| Y - Yager | $r \rightarrow_{Y} s$ | $s^r$ |
| EZ - Early Zadeh | $r \rightarrow_{EZ} s$ | $(r \wedge s) \vee (1 - r)$ |

where $t = 1, \cdots, l^i$ taking partial action frame-by-frame, for the length of an action bounded by $l_{min}$ and $l_{max}$. With these set of implication operators, each inference yields an interval in the range [0, 1]. The upper bound of an inference is given by Eq. (18), and the lower bound is given by Eq. (19). The implication operators must follow the constraint in Eq. (15) for reliable detection.

### 3.4.2. Study on the inference structures

There exists a number of inference structures developed using operators such as $\wedge$, $\vee$ and t-norm [43] that are employed in various applications. For example, the inference structures K7 and K9 delivered good performance for the medical expert system in [22].

$$K_7: \quad R \triangleleft_{K7} S(a,c) = \min\left(\frac{1}{J}\sum_{b \in B}(R(a,b) \rightarrow S(b,c)), \text{OrBot}(\text{AndBot}(R(a,b), S(b,c)))\right) \tag{20}$$

$$K_9: \quad R \triangleleft_{K9} S(a,c) = \min\left(\frac{1}{J}\sum_{b \in B}(R(a,b) \rightarrow S(b,c)), \text{OrBot}(\text{AndTop}(R(a,b), S(b,c)))\right) \tag{21}$$

where $AndTop(p, q) = \min(p, q)$, $AndBot(p, q) = \max(0, p + q - 1)$ and $OrBot(p, q) = \min(1, p + q)$ are the logical connectives. Furthermore, the inference structures instantiated from the original BK subproduct (Eq. (22)) along with the combination of K7 and K9 were applied for scene classification in [21,45].

$$BK: \quad R \triangleleft_{BK} S(a,c) = \frac{1}{J}\sum_{b \in B}(R(a,b) \rightarrow S(b,c)) \tag{22}$$

However, in order to find the suitable inference structure for human action recognition, the detector performance is tested using the classical inference structures: K7, K9 and original BK. The comparison results are shown in Section 4.1.

## 4. Experiments

In order to test the effectiveness of the proposed method, preliminary experiments were performed on the Weizmann human action dataset [55]. Ten natural actions such as: 'run', 'walk', 'skip', 'jack' (jumping-jack), 'jump' (jump-forward-on-two-legs), 'pjump' (jump-in-place-on-two-legs), 'side' (gallop-sideways), 'wave2' (wave-two-hands), 'wave1' (waveone-hand), and 'bend' were performed by nine different people. The preprocessing of images, feature extraction and covariance tracking were performed using the method in [47], further modifying it to generate the part-based human body model with separate tracks for full body, head, torso (arm included) and legs. Sample tracking results are shown in Fig. 7.

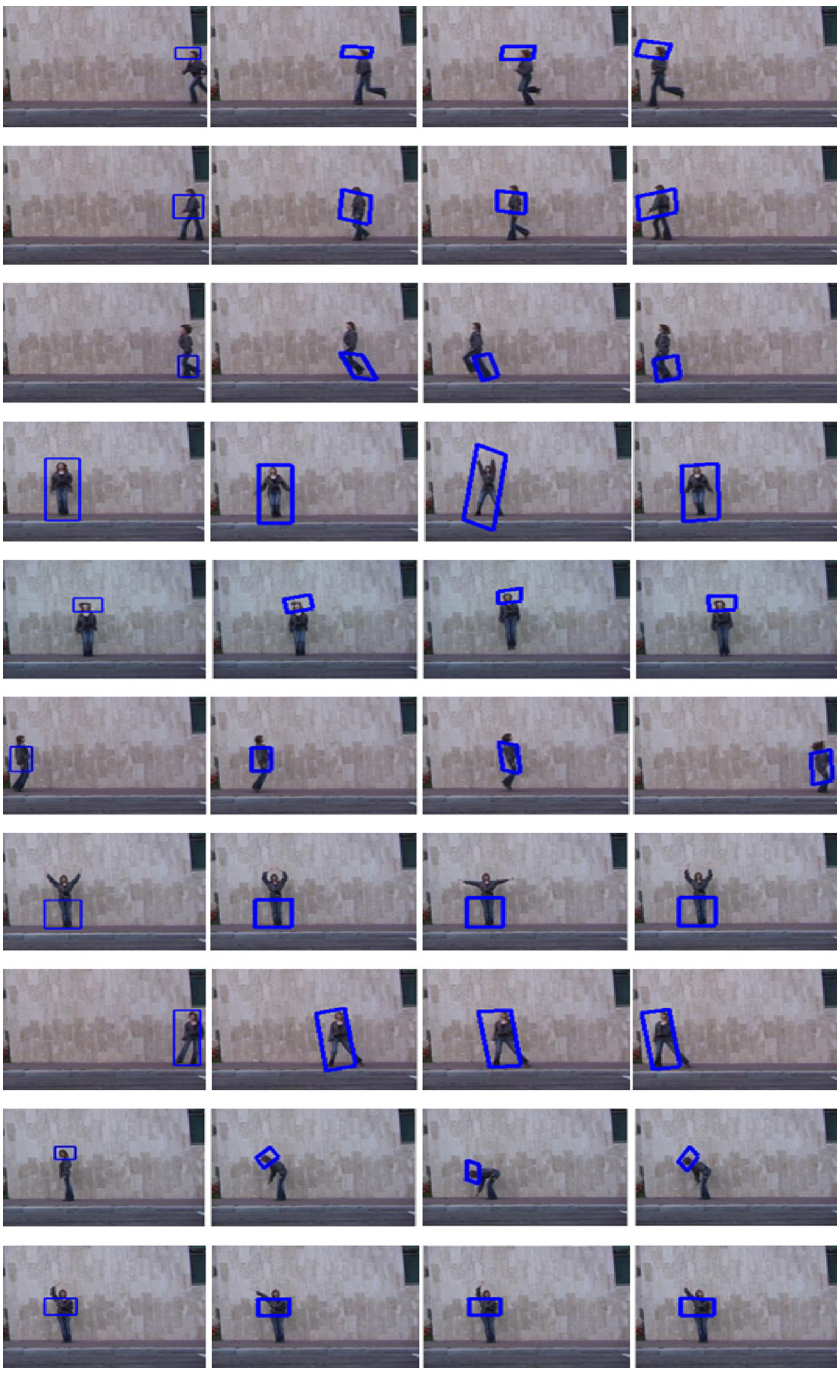
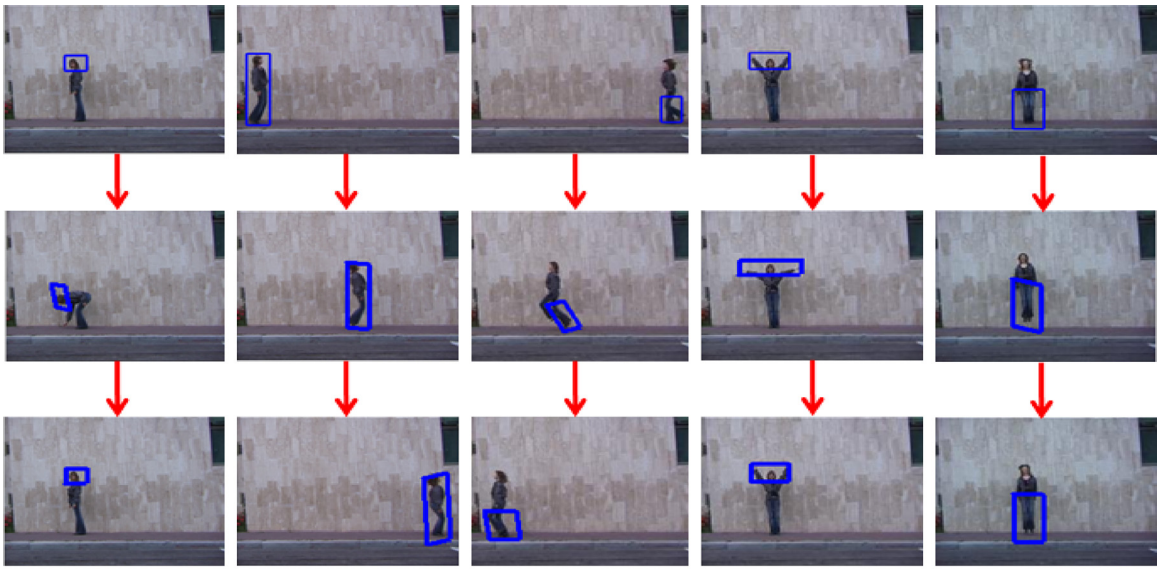


**Fig. 7.** Sample tracking results: From top to bottom row represents the part-based covariance tracking results for run, walk, skip, jack, pjump, jump, wave2, side, bend and wave1 action.

Utilizing the results obtained from Fig. 7, five models were constructed: $m_1$ – model the head movement from start to end frame, $m_2$ – model the position changes of the human body from the first frame, $m_3$ – model the distance between both the legs, $m_4$ – model the hand movement from start to end frame, and $m_5$ – model the distance of the human body from the ground. Fig. 8 presents the model which forms the set $B$ for BK relational product. The membership function $R(f, m)$ is generated by normalizing the results obtained from the model-based covariance tracking using the fuzzy qualitative quantity states ($s = 10$). As can be seen in Table 2, $R(f, m)$ represents the one-to-many relationship between the images (set $A$) and the models (set $B$) describing the degree of belongingness between an image and several models. The membership function $S(m, a)$ represents the relationship between the model (set $B$) and the action (set $C$) being performed. Table 3 highlights the membership function values generated for the one-to-many relationship between model and action, with each model having a degree of belongingness to the action classes. Generating $R$ and $S$ for each image frame, BK subproduct inference engine is invoked. Utilizing Eqs. (16), (18) and (19), human action classification is performed. Since the partial human action is modeled instead of the complete action, the detector is capable of detecting an action early, before its completion.

The proposed detector detects an action when the membership function value exceeds the pre-defined threshold monotonically. It was observed that automatic thresholding doesn't provide optimal



(a) $m_1$ for 'bend' (b) $m_2$ for 'jump' (c) $m_3$ for 'skip' (d) $m_4$ for 'wave2' (e) $m_5$ for 'pjump'

**Fig. 8.** Part-based human body model generated from human motion tracking: $m_1 - m_5$ for five example action sequences.

**Table 2**
Example of membership function $R(f, m)$.

| Frame no. | $m_1$ | $m_2$ | $m_3$ | $m_4$ | $m_5$ |
|---|---|---|---|---|---|
| 1 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 10 | 1.00 | 0.10 | 0.00 | 0.10 | 0.20 |
| 20 | 0.70 | 0.60 | 0.00 | 0.30 | 0.70 |
| 30 | 0.20 | 0.70 | 0.00 | 0.90 | 0.30 |
| 40 | 0.00 | 0.30 | 0.00 | 0.90 | 0.00 |
| 50 | 0.40 | 0.40 | 0.00 | 0.50 | 0.30 |
| 60 | 1.00 | 0.10 | 0.00 | 0.10 | 1.00 |
| 70 | 1.00 | 0.10 | 0.00 | 0.10 | 0.00 |
| 80 | 1.00 | 0.10 | 0.00 | 0.20 | 0.10 |
| 90 | 0.50 | 0.10 | 0.80 | 0.00 | 0.00 |
| 100 | 0.40 | 0.30 | 0.80 | 0.00 | 0.00 |
| 110 | 0.40 | 0.40 | 0.80 | 0.00 | 0.80 |
| 120 | 0.50 | 0.60 | 0.60 | 0.00 | 0.80 |
| 130 | 0.20 | 0.80 | 0.30 | 0.00 | 0.50 |
| 140 | 1.00 | 0.90 | 0.10 | 0.00 | 0.60 |
| 150 | 0.50 | 1.00 | 0.20 | 0.00 | 0.30 |

solution for early detection. It is required to set a fixed threshold value for all the actions in order to detect an action early. In the experiments, results were tested using different threshold values i.e. 0.70, 0.75, 0.80, 0.85 and 0.90. Table 4 presents the results obtained, where 't' refers to the threshold value and the last column represents the percentage of frames observed before the detector triggers the action when the threshold was set to 0.70. It was observed that when the threshold was set to 0.70, the detector is able to detect all the actions performed upon seeing ~23% of the frames on an average. Increasing the threshold to 0.75 misses the detection for only 'run' action, and able to make early detection for all other actions upon seeing ~37% of the frames on average. With the threshold value set at 0.80, the detector successfully detects all actions except 'bend', 'jump', 'run', 'walk' and 'wave1', upon seeing ~60% of the frames on average. Even with the threshold value 0.90, the detector is able to detect 'jack', 'skip' and 'pjump' action upon seeing ~33%, ~23%, and ~52% of the frames respectively. Fig. 9 highlights the experimental results, qualitatively for the ten action classes where the proposed detector detects an action upon observing ~23% of the frames (on average) when the membership function attains a certain threshold (e.g. 0.70 or 0.80) monotonically.

### 4.1. Comparison with the state-of-the-art

The conventional computer vision solutions for early detection of human action includes [9–13,51]. In terms of timeliness and accuracy of detection, MMED proposed in [13,51] outperforms the other algorithms. The experiments were performed on the Auslan dataset (Australian Sign Language), the extended Cohn-Kanade dataset (CK+) and the Weizmann human action dataset. On average, MMED requires seeing ~37% of the sentence for Australian sign language recognition. To detect facial expression (CK+), MMED detects when it completes ~47% of the expression. For human action recognition using Weizmann human action dataset, MMED requires seeing ~40% of the action (with a score of 0.7). In this paper, the experiments were performed using the same human action dataset, and it was found that the detector significantly outperforms MMED where the detector requires seeing ~23% of the image frames on an average in an action video (with membership function score of 0.7). Nonetheless, the computational cost involved in MMED is high as it requires extensive labeling on each of the training samples. Due to the inherent advantages of the BK subproduct inference mechanism, the computational cost involved is lower as compared to MMED, i.e. $O(l)$, where $l$ is length of the action. Moreover, MMED is lacking in in terms of handling the vague feature data and uncertainty involved in the training stage. The proposed method is based on fuzzy BK subproduct and therefore inherits the capabilities of fuzzy theory in handling the uncertainties involved therein using the fuzzy membership function values generated by invoking the BK subproduct inference engine.

However, there exists several methods that employ fuzzy logic for human action recognition. For example, fuzzy inference system was successfully applied in [38,39] for effectively distinguishing the human motion patterns using the flexible membership functions and the fuzzy rules with endurance to the vague feature data. In [41], fuzzy vector quantization incorporated with fuzzy c-means was used to model the human movements with flexibility to support complex continuous actions. In spite of the inherent advantages of fuzzy logic in performing human action recognition, these approaches require seeing the complete action video to detect an action. Hence, these approaches lack in ability to detect an action early and cannot be quantitatively compared with the proposed

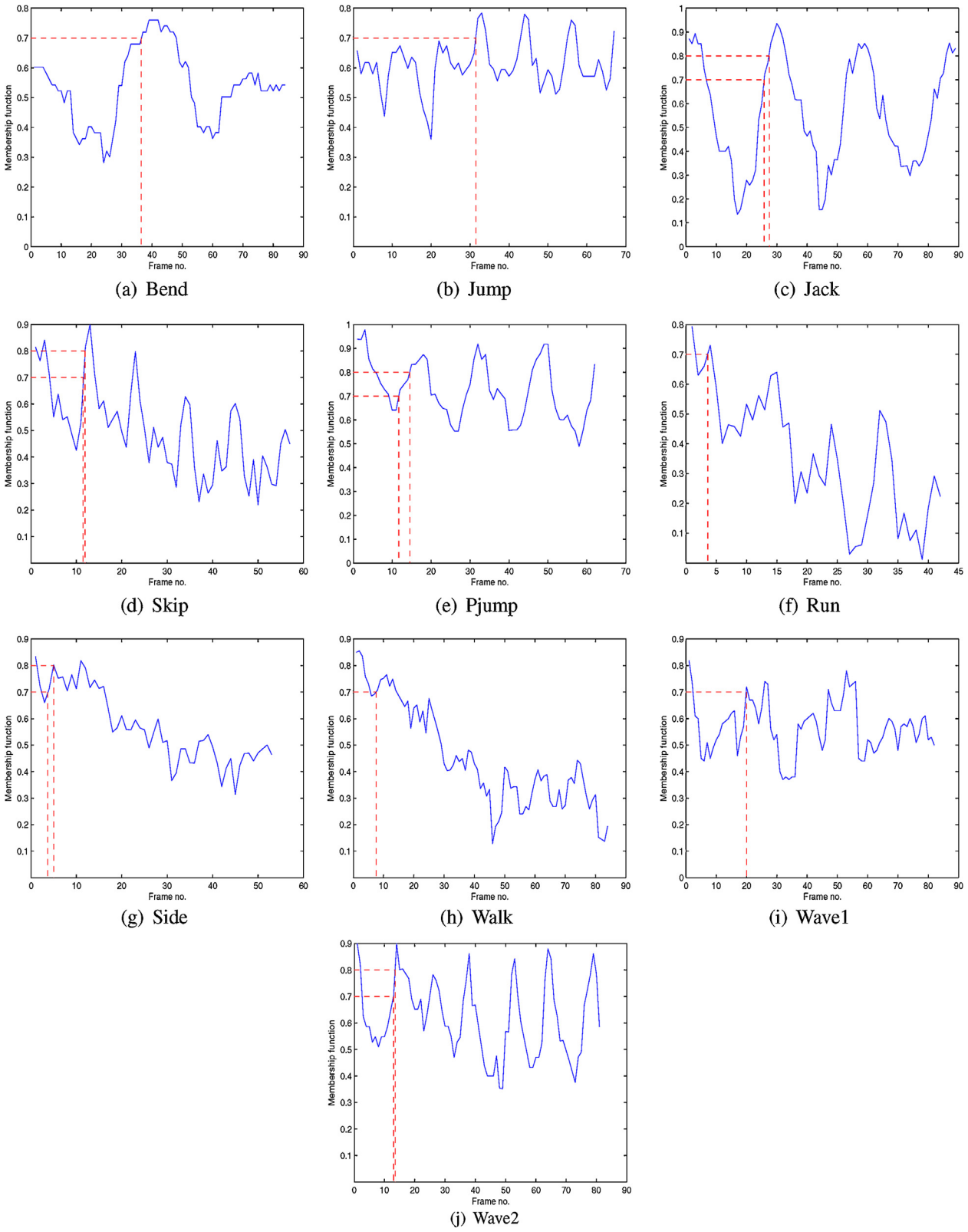



**Fig. 9.** Graphical results for early detection of human action. The detector triggers the action upon seeing ∼23% of the frames on an average when the membership function attains a certain threshold (e.g. 0.70 and 0.80 here) monotonically.

methodology. To the best of the authors' knowledge, no paper has ever employed fuzzy set oriented approaches for early recognition of human action. This claim is supported by a recent survey paper [16]. The method proposed in this paper utilizes the fuzzy BK subproduct inference mechanism for early detection of human action.

Recently, there has been a tremendous growth of research exploring the fusion of intelligent elements using efficient hybrid

**Table 3**
Example of membership function $S(m, a)$.

| Model | Bend | Jump | Jack | Skip | Pjump | Run | Side | Walk | Wave1 | Wave2 |
|---|---|---|---|---|---|---|---|---|---|---|
| $m_1$ | 0.60 | 0.80 | 0.01 | 0.80 | 0.11 | 0.80 | 0.70 | 0.80 | 0.00 | 0.00 |
| $m_2$ | 0.01 | 0.80 | 0.12 | 0.90 | 0.20 | 0.90 | 0.85 | 0.90 | 0.00 | 0.00 |
| $m_3$ | 0.01 | 0.10 | 0.82 | 0.25 | 0.01 | 0.85 | 0.75 | 0.88 | 0.00 | 0.00 |
| $m_4$ | 0.70 | 0.01 | 0.90 | 0.15 | 0.18 | 0.60 | 0.65 | 0.60 | 0.50 | 0.90 |
| $m_5$ | 0.01 | 0.25 | 0.20 | 0.20 | 0.30 | 0.20 | 0.01 | 0.01 | 0.00 | 0.00 |

techniques. For example, [14,15] effectively integrated fuzzy logic with machine learning techniques for human action recognition where optimum membership function and flexible fuzzy rules were used to infer human behavior. However, the conventional hybrid methods for human action recognition are not capable of inferring an action early, i.e. before its completion. This paper reveals the inherent strength of hybridization of computational methods (computer vision solutions as well as fuzzy BK subproduct) for early human action detection in a way that the strength of fuzzy set theory can alleviate the limitation of computer vision solutions, where partial human actions were modeled for early detection, instead of a complete action. To the very best of the authors' knowledge, this is the first work in the community that employs hybrid technique for solving the problem of early human action detection and it stands out against other conventional methods with good detection rate where the detector requires seeing only ∼23% of the frames on average to detect an action.

In order to justify the choice of employing BK subproduct for human action recognition, the performance of the detector was evaluated using the classical inference structures: K7, K9

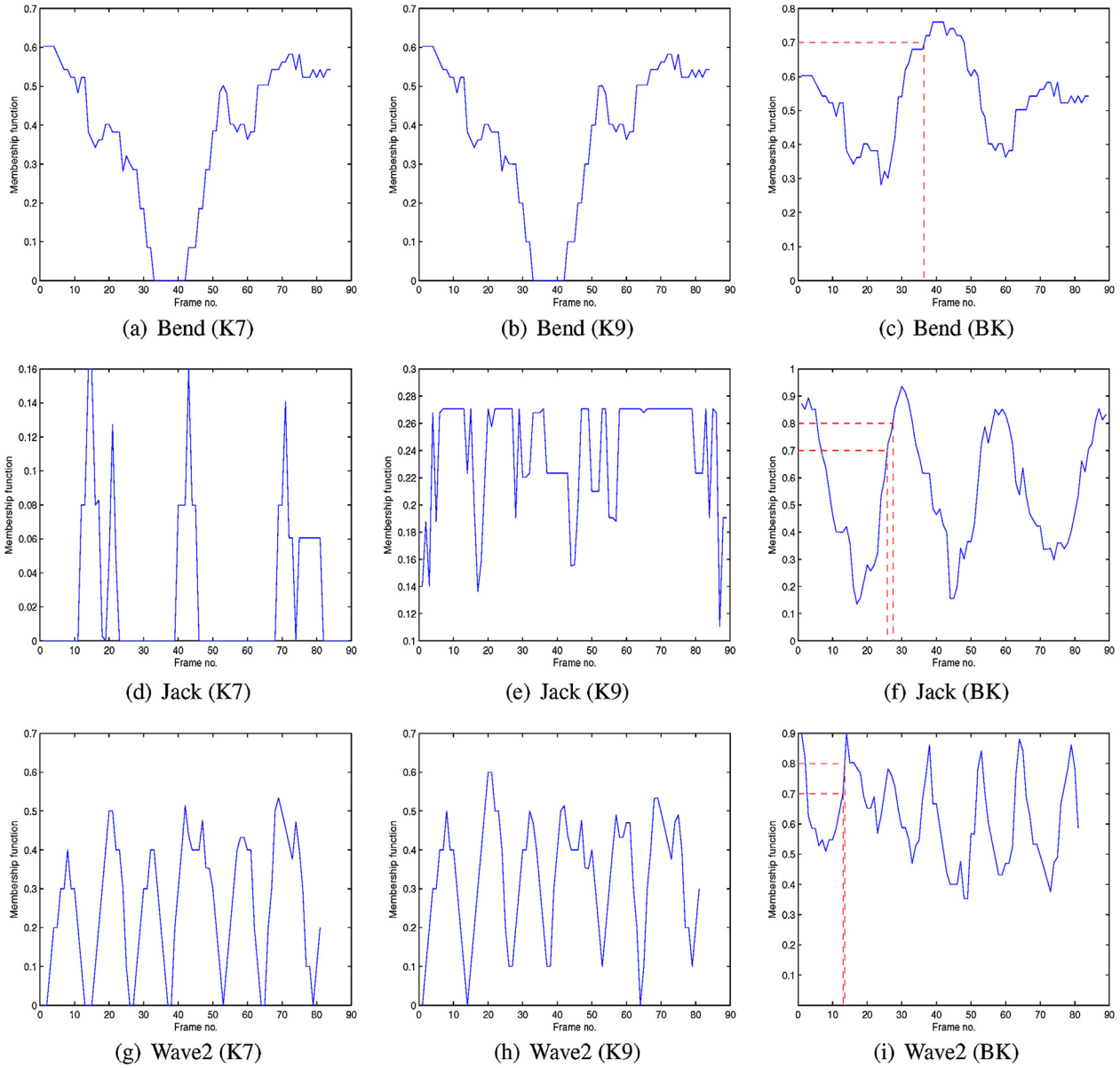


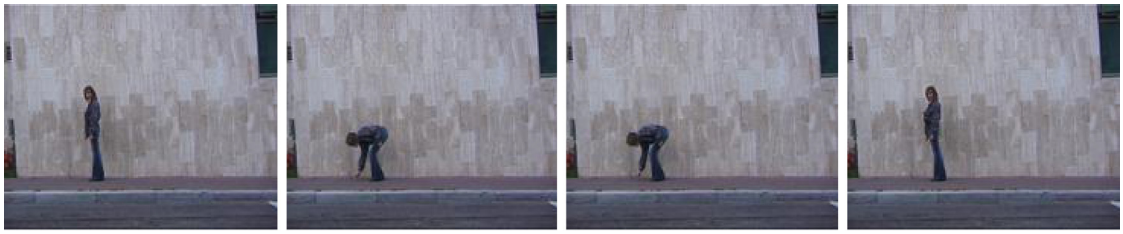

**Fig. 10.** Graphical results representing the detector performance using K7, K9 and Original BK inference structure (BK) for three example action: bend, jack and wave2 (wave-two-hands).

**Table 4**
Results for early detection of human action.

| Action | Total no. of frames | $t = 0.70$ | $t = 0.75$ | $t = 0.80$ | $t = 0.85$ | $t = 0.90$ | Frames seen (%) |
|---|---|---|---|---|---|---|---|
| Bend | 84 | 36 | 39 | – | – | – | 42.85 |
| Jump | 67 | 31 | 32 | – | – | – | 46.26 |
| Jack | 89 | 25 | 26 | 27 | 28 | 29 | 28.08 |
| Skip | 57 | 11 | 12 | 12 | 13 | 13 | 19.29 |
| Pjump | 62 | 12 | 13 | 15 | 17 | 32 | 19.35 |
| Run | 42 | 4 | – | – | – | – | 9.52 |
| Side | 53 | 4 | 5 | 5 | – | – | 7.54 |
| Walk | 84 | 8 | 9 | – | – | – | 9.52 |
| Wave1 | 82 | 20 | 52 | – | – | – | 24.39 |
| Wave2 | 81 | 13 | 13 | 14 | 14 | – | 16.04 |

**Table 5**
Membership function values for inference structures.

| Inference structure | Frame no. | Bend | Jump | Jack | Skip | Pjump | Run | Side | Walk | Wave1 | Wave2 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| K7 | 1 | 0.60 | 0.66 | 0.00 | 0.00 | 0.00 | 0.12 | 0.00 | 0.00 | 0.00 | 0.00 |
| | 10 | 0.52 | 0.50 | 0.00 | 0.39 | 0.00 | 0.42 | 0.00 | 0.00 | 0.50 | 0.30 |
| | 20 | 0.40 | 0.36 | 0.05 | 0.10 | 0.00 | 0.23 | 0.00 | 0.00 | 0.00 | 0.50 |
| | 30 | 0.19 | 0.50 | 0.00 | 0.37 | 0.00 | 0.16 | 0.03 | 0.03 | 0.00 | 0.30 |
| | 40 | 0.00 | 0.50 | 0.08 | 0.10 | 0.02 | 0.06 | 0.33 | 0.00 | 0.30 | 0.30 |
| K9 | 1 | 0.60 | 0.66 | 0.14 | 0.55 | 0.02 | 0.61 | 0.29 | 0.13 | 0.00 | 0.00 |
| | 10 | 0.52 | 0.60 | 0.27 | 0.43 | 0.17 | 0.53 | 0.62 | 0.75 | 0.50 | 0.40 |
| | 20 | 0.40 | 0.36 | 0.27 | 0.50 | 0.17 | 0.23 | 0.55 | 0.64 | 0.00 | 0.60 |
| | 30 | 0.20 | 0.60 | 0.22 | 0.38 | 0.17 | 0.16 | 0.52 | 0.43 | 0.40 | 0.40 |
| | 40 | 0.00 | 0.57 | 0.22 | 0.29 | 0.17 | 0.18 | 0.49 | 0.43 | 0.50 | 0.40 |
| Original BK | 1 | 0.60 | 0.66 | 0.87 | 0.81 | 0.94 | 0.79 | 0.83 | 0.85 | 0.82 | 0.90 |
| | 10 | 0.52 | 0.65 | 0.47 | 0.43 | 0.64 | 0.53 | 0.71 | 0.75 | 0.52 | 0.55 |
| | 20 | 0.40 | 0.36 | 0.28 | 0.50 | 0.70 | 0.23 | 0.61 | 0.64 | 0.72 | 0.65 |
| | 30 | 0.54 | 0.61 | 0.94 | 0.38 | 0.75 | 0.16 | 0.52 | 0.43 | 0.54 | 0.59 |
| | 40 | 0.76 | 0.57 | 0.46 | 0.29 | 0.56 | 0.18 | 0.49 | 0.43 | 0.60 | 0.67 |



(a) 0.00 (b) 0.42 (c) 0.46 (d) 1.00

**Fig. 11.** NTtoD for *bend*. (a) Onset frame, (b) NTtoD with threshold 0.70 (our detector fires), (c) NTtoD with threshold 0.80, (d) Peak frame.

and Original BK subproduct (BK). It was found that overall the original BK performed best for all the action classes as shown in Table 5. Whereas, K9 delivered comparable results for some action sequences (e.g. bend, jump, skip, run and walk) and K7 performed fairly poorer for all the action classes. Fig. 10 evaluates the results qualitatively for the three example action classes: bend, jack and wave2. It can be observed from the graphical representation that the membership function values generated using K7 and K9 inference structures are much lower as compared to original BK. Therefore, it is deduced that original BK subproduct is the most suitable inference structure to be used in the application under consideration.

To evaluate the timeliness of detection, NTtoD (Normalized Time to Detect) was used. Assume for a given action sequence, where the action occurs from the start frame $s$ to end frame $e$, the detector triggers the action at time $t$. For a successful detection, $s \leq t \leq e$, NTtoD is defined as $\frac{t-s+1}{e-s+1}$ i.e. the fraction of action occurred. When $t < s$, NTtoD=0 i.e. false detection, and when $t > e$, NTtoD = $\infty$ i.e. false rejection. For the well-known classifiers (e.g. SVM, KNN), the classification is performed by observing the complete action sequence and therefore NTtoD for SVM and KNN is 1. NTtoD for the detector in this work is as follows: bend = 0.42, jump = 0.46, jack = 0.28, skip = 0.19, pjump = 0.19, run = 0.09, side = 0.07, walk = 0.09, wave1 = 0.24 and wave2 = 0.16. Fig. 11 highlights the NTtoD results obtained using the detector for bend action.

## 5. Conclusion and future research

This paper takes the initiative to fuse the benefits of both computer vision and fuzzy set theory to develop a hybrid technique capable of performing human action recognition early. Human action classification problem is modified into frame-by-frame level classification where the partial human actions were modeled to enable early detection. The membership function values generated for each human action are utilized to infer an action. Detection is triggered when the membership function attains a pre-defined threshold monotonically. The experimental results demonstrate the capability of the proposed detector to carry out reliable early human action detection. On average, the detector is able to infer an action upon viewing ~23% of the frames for test data under the experimental settings. For future work, the plan is to introduce and perform experiments on fuzzy dataset i.e. dataset with fuzzy ground truths to overcome the limitation of the current datasets

being mutually exclusive, allowing a data to belong to one action class only at a time. In view of the encouraging results obtained in this work, hybridization of deep learning and fuzzy set theory for human action recognition can also be explored as a future work.

## Acknowledgments

This research is supported by the High Impact MoE Grant UM.C/625/1/HIR/MoE/FCSIT/08, H-22001-00-B00008 from the Ministry of Education Malaysia. The authors would like to thank Chee Kau Lim for the useful discussion on the BK Sub-triangle product.

## References

[1] Y. Hatakeyama, A. Mitsuta, K. Hirota, Detection algorithm for color dynamic images by multiple surveillance cameras under low luminance conditions based on fuzzy corresponding map, Appl. Soft Comput. 8 (4) (2008) 1344–1353.
[2] O.P. Popoola, K. Wang, Video-based abnormal human behavior recognition: a review, IEEE Trans. Syst. Man Cybern. C: Appl. Rev. 42 (6) (2012) 865–878.
[3] I.S. Kim, H.S. Choi, K.M. Yi, J.Y. Choi, S.G. Kong, Intelligent visual surveillance-a survey, Int. J. Control Autom. Syst. 8 (5) (2010) 926–939.
[4] D. Anderson, J.M. Keller, M. Skubic, X. Chen, Z. He, Recognizing falls from silhouettes, in: EMBS, 2006, pp. 6388–6391.
[5] D. Sanchez-Valdes, A. Alvarez-Alvarez, G. Trivino, Walking pattern classification using a granular linguistic analysis, Appl. Soft Comput. 33 (2015) 100–113.
[6] D. Anderson, R.H. Luke, J.M. Keller, M. Skubic, M.J. Rantz, M.A. Aud, Modeling human activity from voxel person using fuzzy logic, IEEE Trans. Fuzzy Syst. 17 (1) (2009) 39–49.
[7] M.D. Rodriguez, J. Ahmed, M. Shah, Action mach: a spatio-temporal maximum average correlation height filter for action recognition, in: CVPR, 2008, pp. 1–8.
[8] E. Yeguas-Bolivar, R. Mu noz-Salinas, R. Medina-Carnicer, A. Carmona-Poyato, Comparing evolutionary algorithms and particle filters for markerless human motion capture, Appl. Soft Comput. 17 (2014) 153–166.
[9] M. Ryoo, Human activity prediction: early recognition of ongoing activities from streaming videos, in: ICCV, 2011, pp. 1036–1043.
[10] G. Yu, J. Yuan, Z. Liu, Predicting human activities using spatio-temporal structure of interest points, in: ACM-MM, 2012, pp. 1049–1052.
[11] M. Ryoo, T. J. Fuchs, L. Xia, J. Aggarwal, L. Matthies, Early recognition of human activities from first-person videos using onset representations, arXiv preprint arXiv:1406.5309.
[12] K. Li, Y. Fu, Arma-hmm: a new approach for early recognition of human activity, in: ICPR, 2012, pp. 1779–1782.
[13] M. Hoai, F. De la Torre, Max-margin early event detectors, in: CVPR, 2012, pp. 2863–2870.
[14] G. Acampora, P. Foggia, A. Saggese, M. Vento, Combining neural networks and fuzzy systems for human behavior understanding, in: AVSS, 2012, pp. 88–93.
[15] M.-S. Hosseini, A.-M. Eftekhari-Moghadam, Fuzzy rule-based reasoning approach for event detection and annotation of broadcast soccer video, Appl. Soft Comput. 13 (2) (2013) 846–866.
[16] C.H. Lim, E. Vats, C.S. Chan, Fuzzy human motion analysis: a review, Pattern Recogn. 48 (5) (2015) 1773–1796.
[17] W. Bandler, L. Kohout, Fuzzy power sets and fuzzy implication operators, Fuzzy Sets Syst. 4 (1) (1980) 13–30.
[18] C.K. Lim, C.S. Chan, A weighted inference engine based on interval-valued fuzzy relational theory, Expert Syst. Appl. 42 (7) (2015) 3410–3419.
[19] L.-D. Bui, Y.-G. Kim, An obstacle-avoidance technique for autonomous underwater vehicles based on BK-products of fuzzy relation, Fuzzy Sets Syst. 157 (4) (2006) 560–577.
[20] R. Groenemans, E. Van Ranst, E. Kerre, Fuzzy relational calculus in land evaluation, Geoderma 77 (2) (1997) 283–298.
[21] E. Vats, C.K. Lim, C.S. Chan, A bk subproduct approach for scene classification, in: IEEJ IEVC, 2012, pp. 1–5.
[22] C.K. Lim, C.S. Chan, Logical connectives and operativeness of bk sub-triangle product in fuzzy inferencing, Int. J. Fuzzy Syst. 13 (4) (2011) 237–245.
[23] H. Liu, G.M. Coghill, Fuzzy qualitative trigonometry, in: IEEE Conference on Systems, Man and Cybernetics, Vol. 2, 2005, pp. 1291–1296.
[24] C.S. Chan, H. Liu, Fuzzy qualitative human motion analysis, IEEE Trans. Fuzzy Syst. 17 (4) (2009) 851–862.
[25] E. Vats, C.S. Chan, Early human actions detection using BK sub-triangle product, in: FUZZ-IEEE, 2015.
[26] J.K. Aggarwal, Q. Cai, W. Liao, B. Sabata, Articulated and elastic non-rigid motion: a review, in: Motion of Non-Rigid and Articulated Objects Workshop, 1994, pp. 2–14.
[27] C. Cédras, M. Shah, Motion-based recognition a survey, Image Vis. Comput. 13 (2) (1995) 129–155.
[28] J.K. Aggarwal, Q. Cai, Human motion analysis: a review, in: Nonrigid and Articulated Motion Workshop, 1997, pp. 90–102.
[29] D.M. Gavrila, The visual analysis of human movement: a survey, Comput. Vis. Image Underst. 73 (1) (1999) 82–98.
[30] O. Duchenne, I. Laptev, J. Sivic, F. Bach, J. Ponce, Automatic annotation of human actions in video, in: ICCV, 2009, pp. 1491–1498.
[31] A. Patron-Perez, M. Marszalek, A. Zisserman, I.D. Reid, High five: recognising human interactions in TV shows, in: BMVC, Vol. 1, Citeseer, 2010, p. 2.
[32] M. Hoai, Z.-Z. Lan, F. De la Torre, Joint segmentation and classification of human actions in video, in: CVPR, 2011, pp. 3265–3272.
[33] K.K. Reddy, M. Shah, Recognizing 50 human action categories of web videos, Mach. Vis. Appl. 24 (5) (2013) 971–981.
[34] W. Brendel, S. Todorovic, Learning spatiotemporal graphs of human activities, in: ICCV, 2011, pp. 778–785.
[35] M.H. Nguyen, L. Torresani, F. De la Torre, C. Rother, Weakly supervised discriminative localization and classification: a joint learning process, in: ICCV, 2009, pp. 1925–1932.
[36] M.S. Ryoo, J.K. Aggarwal, Semantic representation and recognition of continued and recursive human activities, Int. J. Comput. Vis. 82 (1) (2009) 1–24.
[37] S.D. Tran, L.S. Davis, Event modeling and recognition using Markov logic networks, in: Computer Vision-ECCV 2008, Springer, 2008, pp. 610–623.
[38] J.-M. Le Yaouanc, J.-P. Poli, A fuzzy spatio-temporal-based approach for activity recognition, in: Advances in Conceptual Modeling, 2012, pp. 314–323.
[39] B. Yao, H. Hagras, M.J. Alhaddad, D. Alghazzawi, A fuzzy logic-based system for the automation of human behavior recognition using machine vision in intelligent environments, Soft Comput. (2014) 1–8.
[40] K. Mozafari, N.M. Charkari, H.S. Boroujeni, M. Behrouzifar, A novel fuzzy hmm approach for human action recognition in video, in: Knowledge Technology, Springer, 2012, pp. 184–193.
[41] N. Gkalelis, A. Tefas, I. Pitas, Combining fuzzy vector quantization with linear discriminant analysis for continuous human movement recognition, IEEE Trans. Circuits Syst. Video Technol. 18 (11) (2008) 1511–1521.
[42] C.S. Chan, H. Liu, W.K. Lai, Fuzzy qualitative complex actions recognition, in: FUZZ-IEEE, 2010, pp. 1–8.
[43] L. Kohout, Interval-based reasoning in medical diagnosis, in: IIS, 1997, pp. 32-32.
[44] L.J. Kohout, W. Bandler, Relational-product architectures for information processing, Inf. Sci. 37 (1) (1985) 25–37.
[45] E. Vats, C.K. Lim, C.S. Chan, An improved BK sub-triangle product approach for scene classification, J. Intell. Fuzzy Syst. (2015) 1–9 (Preprint).
[46] W. Bandler, L.J. Kohout, Semantics of implication operators and fuzzy relational products, Int. J. Man Mach. Stud. 12 (1) (1980) 89–116.
[47] F. Porikli, O. Tuzel, P. Meer, Covariance tracking using model update based on lie algebra, in: CVPR, Vol. 1, 2006, pp. 728–735.
[48] M. Cristani, R. Raghavendra, A. Del Bue, V. Murino, Human behavior analysis in video surveillance: a social signal processing perspective, Neurocomputing 100 (2013) 86–97.
[49] Y. Cao, D. Barrett, A. Barbu, S. Narayanaswamy, H. Yu, A. Michaux, Y. Lin, S. Dickinson, J.M. Siskind, S. Wang, Recognize human activities from partially observed videos, in: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2013, pp. 2658–2665.
[50] Y. Kong, D. Kit, Y. Fu, A discriminative model with multiple temporal scales for action prediction, in: Computer Vision-ECCV 2014, Springer, 2014, pp. 596–611.
[51] M. Hoai, F. De la Torre, Max-margin early event detectors, Int. J. Comput. Vis. 107 (2) (2014) 191–202.
[52] I. Tsochantaridis, T. Joachims, T. Hofmann, Y. Altun, Large margin methods for structured and interdependent output variables, J. Mach. Learn. Res. (2005) 1453–1484.
[53] C.R. Wren, A. Azarbayejani, T. Darrell, A.P. Pentland, Pfinder, Real-time tracking of the human body, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (1997) 780–785.
[54] A.M.M.S. Ullah, D. D'Addona, N. Arai, DNA based computing for understanding complex shapes, Biosystems 117 (2014) 40–53.
[55] L. Gorelick, M. Blank, E. Shechtman, M. Irani, R. Basri, Actions as space-time shapes, IEEE Trans. Pattern Anal. Mach. Intell. 29 (12) (2007) 2247–2253.