Computer Vision and Image Understanding 113 (2009) 80–89

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Linguistic summarization of video for fall detection using voxel person and fuzzy logic

Derek Anderson [a,*], Robert H. Luke [a,1], James M. Keller [a,1], Marjorie Skubic [a,1], Marilyn Rantz [b,2], Myra Aud [b,2]

[a] Department of Electrical and Computer Engineering, University of Missouri, 349 Engineering Building West, Columbia, MO 65211-2300, USA
[b] Sinclair School of Nursing, University of Missouri, Columbia, MO 65211, USA

ARTICLE INFO

Article history:
Received 7 November 2007
Accepted 11 July 2008
Available online 29 July 2008

Keywords:
Linguistic summarization
Activity analysis
Fuzzy logic
Fall detection
Eldercare
Voxel person

ABSTRACT

In this paper, we present a method for recognizing human activity from linguistic summarizations of temporal fuzzy inference curves representing the states of a three-dimensional object called voxel person. A hierarchy of fuzzy logic is used, where the output from each level is summarized and fed into the next level. We present a two level model for fall detection. The first level infers the states of the person at each image. The second level operates on linguistic summarizations of voxel person's states and inference regarding activity is performed. The rules used for fall detection were designed under the supervision of nurses to ensure that they reflect the manner in which elders perform these activities. The proposed framework is extremely flexible. Rules can be modified, added, or removed, allowing for per-resident customization based on knowledge about their cognitive and physical ability.

1. Introduction

Falls are a severe problem among the elderly. Many elders fall and sustain an injury or remain on the floor for long durations until someone discovers them, further compounding the severity. Our goal is the continuous monitoring of human activity for the assessment of the "well-being" of a resident and the detection of abnormal or dangerous events, such as falls. It is important that both theories and realistic technologies be developed for recognizing elderly activity, and that they do so in a non-invasive fashion. Video sensors are a rich source of information that can be used to monitor a scene, but privacy is always a concern. To preserve privacy, segmentation of the human from an image results in a silhouette, a binary map that distinguishes the individual from the background. The raw video is not stored and only silhouettes are used to track the individual's activity.

Silhouette extraction, namely, segmenting the human from an image with the camera at a fixed location, is the first stage in video-based activity analysis. The standard approach involves constructing a background model and regions in subsequent images with significantly different characteristics are classified as foreground [1–10]. Stauffer and Grimson introduced an adaptive method for background modeling and subtraction that utilizes a mixture of Gaussians per pixel with a real-time, online approximation to the model update [1]. Oliver et al. carry out foreground segmentation in eigenspace, where the background is modeled as an eigenbackground, however, no model update method was proposed [2]. These two well known approaches focus on adaptation and background modeling at a relatively low level of computer vision. They do not present robust features for change detection or medium to high level computer vision algorithms for region identification and tracking. The Wallflower algorithm addresses a wider range of extreme real-world conditions for complex and dynamic environments through pixel-level probabilistic background prediction with a Wiener filter, region-level processing, and heuristics for global sudden change detection and model correction [3]. We proposed an adaptive system that uses higher level computer vision for background modeling and reliable change detection through fusing new texture and color histogram-based descriptors and a modified hue, saturation, and value (HSV) space for shadow removal [4].

While change detection is full of technical challenges, even after the human is segmented from the background a larger problem arises regarding the higher level processing of this information for recognizing activity and detecting deviations from patterns of normal activity. The most widely accepted approaches to modeling human activity include; graphical models [11], dynamic Bayesian networks [12], also known as dynamic graphical models, and more specifically, hidden Markov models (HMMs) [2,13] and its variants

* Corresponding author. Fax: +1 573 882 0397.
E-mail addresses: dtaxtd@mizzou.edu (D. Anderson), rhl3db@mizzou.edu (R.H. Luke), kellerj@missouri.edu (J.M. Keller), skubicm@missouri.edu (M. Skubic), rantzm@health.missouri.edu (M. Rantz), audm@health.missouri.edu (M. Aud).
[1] Fax: +1 573 882 0397.
[2] Fax: +1 573 884 4544.

1077-3142/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2008.07.006

(hierarchical HMMs [14], entropic-HMMs [15], coupled-HMMs [2,16], etc).

In the area of short-term activity recognition, we used HMMs for fall detection [13]. Our preliminary results indicate that a single camera, geometric features calculated from silhouettes, and HMMs can be used to detect some types of falls under a constrained set of view dependent assumptions about how and where activities are performed in the environment. However, while HMMs can be used to identify a maximum likely model, from one of K known models, they are not presently sufficient for rejecting unknown activity. Thome and Miguet used hierarchical HMMs (HHMM) for video-based fall detection [14]. The interesting aspect of that research is the feature employed in the model. They use image rectification to derive relationships between the three-dimensional angle corresponding to the individual's major orientation and the principal axis of an ellipse fit to the human in a two-dimensional image. The HHMM is hand designed and operates on an observation sequence of rectified angles.

Martin et al. presented a soft-computing approach to monitoring the "well-being" of elders over long time periods using non-video sensors such as passive infrared, toggle switches, vibration, temperature, and pressure sensors [17]. They outlined data analysis methods based on fuzzy reasoning, statistics, association analysis, and trend analysis. Procedures for interpreting firings from relatively simple sensors into fuzzy summaries were presented. These summaries assist in characterizing resident's trends and aid in answering queries about deviations from these patterns, such as "has the occupant's sleep pattern changed significantly in the past few months".

In [18], we presented a method for constructing a three-dimensional representation of the human from silhouettes acquired from multiple cameras monitoring the same scene. Fuzzy logic is used to determine the membership degree of the person to a pre-determined number of states at each image. In this paper, a method is presented for generating a significantly smaller number of rich linguistic summaries of the human's state over time, in comparison to the large number of state decisions made at each image, and a procedure is introduced for inferring activity from features calculated from linguistic summarizations. Summarization and activity inference makes fall detection possible, something that was not accomplished in our earlier work. The next section is an overview of voxel person construction and state reasoning. Some material is reported again here because it is necessary for understanding the approach taken in this paper.

2. Fuzzy logic for voxel person state classification

Our approach to monitoring human activity is based on fuzzy set theory. Fuzzy set theory, introduced by Lotfi A. Zadeh in 1965 [19], is an extension of classical set theory. The memberships of elements in a set are allowed to vary in their degree, instead of being restricted to two values, as in classical set theory. A fuzzy set is defined over a particular domain, and it is characterized by a membership function that maps elements from the domain to a real valued number, $\mu_A : A \rightarrow [0,1]$. The fuzzy sets used in this paper are trapezoidal membership functions, which are characterized according to four ordered numbers, {a,b,c,d}. The membership of the element x in the fuzzy set A is

$$\mu_A(x) = \text{maximum}\left(\text{minimum}\left(\frac{(x-a)}{(b-a)}, 1, \frac{(d-x)}{(d-c)}\right), 0\right).$$

The way in which fuzzy set theory includes and models uncertainty has led to extremely valuable applications in mathematics and engineering [20–23]. One of the more well known branches of fuzzy set theory is fuzzy logic, introduced by Zadeh in 1973 [20]. Fuzzy logic is a powerful framework for performing automated reasoning. An inference engine operates on rules that are structured in an IF–THEN format. The IF part of the rule is called the antecedent, while the THEN part of the rule is called the consequent. Rules are constructed from linguistic variables. These variables take on the fuzzy values or fuzzy terms that are represented as words and modeled as fuzzy subsets of an appropriate domain. An example is the fuzzy linguistic variable "height of voxel person's centroid", which can assume the terms low, medium, and high, all defined as membership functions over an appropriate numerical domain. In this work, we use the standard Mamdani-Assilion fuzzy inference system [20,24].

What is needed in the area of human activity analysis is not another non-interpretable likelihood value that is useful for classifying one of K known models or the ad hoc training of garbage models [25] for reducing false alarms, but a confidence value that can be understood and reliably used to reject a wide range of unknown activities. The core representation and computing basis in our work is significantly different from most. We believe that fuzzy set theory and fuzzy logic are necessary in order to address the inherent uncertainty related to modeling and inferring human activity. Linguistic variables are used to describe features extracted from a three-dimensional representation of the human. A separate set of linguistic variables are used for representing the human's state and activity. Fuzzy logic is used for inferring the state and activity. The system's output are membership values that reside in [0,1], fuzzy sets (terms) have been defined and assist in the interpretation of these values, and fuzzy logic is the inference mechanism. A fuzzy approach also has the advantage that the rules and linguistic variables are understandable and simplify addition, removal, and modification of the system's knowledge.

Multiple cameras that jointly view the same environment are crucial for the reliable recognition of activity. Different viewpoints assist in coping with issues like occlusion and makes the construction of three-dimensional objects possible. After silhouettes are individually extracted from each camera in a scene, a three-dimensional representation of the human is constructed in voxel space, which we call voxel person. Like pixels in a two-dimensional image, a voxel (volume element) is an element resulting from a discretization of three-dimensional space. A voxel is defined here as a non-overlapping cube. The set of voxels belonging to voxel person at time t are $V_t = \{\vec{v}_{t,1}, \vec{v}_{t,2}, \ldots, \vec{v}_{t,P}\}$, where the center of the jth voxel at time t is $\vec{v}_{t,j} = \langle x_j, y_j, z_j \rangle^{\mathrm{T}}$. The capture time for each camera is recorded and the silhouettes, one from each camera, that are the closest in time are used to build V_t. The construction of voxel person from a single camera is the planar extension of the silhouette along the direction of the camera viewing angle. Voxels in the monitored space that are intersected by this planar extension are identified. The projection procedure involves using the camera's intrinsic parameters to estimate pixel rays, and these rays are tested for intersection with voxels [18]. Voxel person, according to camera i ($1 \leqslant i \leqslant C$) at time t is V_t^i, whose cardinality, $|V_t^i|$, is P_i. The planar extensions of voxel person from multiple cameras, $\{V_t^1, \ldots, V_t^C\}$, are combined using an operation, such as intersection, $V_t = \Lambda_{i=1}^{c} V_t^i$, to assemble a more accurate object representation. An illustration of voxel person construction from two cameras is shown in Fig. 1. In [18], further processing is performed on voxel person to remove additional shadows and reflections given a priori knowledge about the three-dimensional environment. Voxel person's volume is analyzed to detect error time intervals, and an efficient method for dynamically increasing the resolution (detail) of voxel person is discussed.

For each image, the goal is the calculation of the membership degree of voxel person to a set of pre-determined states. This state information is used to infer activity. An activity is characterized according to state duration, frequency of state visitation, and state