Expert Systems with Applications 41 (2014) 574–587

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Fast action recognition using negative space features

Shah Atiqur Rahman [a,*], Insu Song [a], M.K.H. Leung [b], Ickjai Lee [c], Kyungmi Lee [c]

[a] School of Business and IT, James Cook University Australia, Singapore 574421, Singapore
[b] FICT, Universiti Tunku Abdul Rahman (Kampar), Malaysia
[c] School of Business (IT), James Cook University Cairns, QLD 4870, Australia

ARTICLE INFO

Keywords:
Action recognition
Negative space action descriptors
Silhouette
Fuzzy membership
Implicit method
Cycle length
Fish actions

ABSTRACT

Due to the number of potential applications and their inherent complexity, automatic capture and analysis of actions have become an active research area. In this paper, an implicit method for recognizing actions in a video is proposed. Existing implicit methods work on the regions of subjects, but our proposed system works on the surrounding regions, called negative spaces, of the subjects. Extracting features from negative spaces facilitates the system to extract simple, yet effective features for describing actions. These negative-space based features are robust to deformed actions, such as complex boundary variations, partial occlusions, non-rigid deformations and small shadows. Unlike other implicit methods, our method does not require dimensionality reduction, thereby significantly improving the processing time. Further, we propose a new method to detect cycles of different actions automatically. In the proposed system, first, the input image sequence is background segmented and shadows are eliminated from the segmented images. Next, motion based features are computed for the sequence. Then, the negative space based description of each pose is obtained and the action descriptor is formed by combining the pose descriptors. Nearest Neighbor classifier is applied to recognize the action of the input sequence. The proposed system was evaluated on both publically available action datasets and a new fish action dataset for comparison, and showed improvement in both its accuracy and processing time. Moreover, the proposed system showed very good accuracy for corrupted image sequences, particularly in the case of noisy segmentation, and lower frame rate. Further, it has achieved highest accuracy with lowest processing time compared with the state-of-art methods.

1. Introduction

One of the general goals of artificial intelligence is to design machines which act intelligently and in a human-like manner. For a machine to be truly intelligent and useful, it must have the ability to perceive the environment in which it is embedded. It needs to be able to extract information from its environment independently, rather than relying on the information supplied by users externally. The visual analysis of human motion attempts to detect, track and identify people, and more generally, to interpret behaviors of subjects from sequences of images, in which humans perform certain actions. Such technologies will allow us to communicate with machines more easily than before, by allowing more advanced input modalities, such as gestures. For example, people extend their hands to robots to do a handshake. In reply, the robots may do the same to make the handshake happen by recognizing the action of the people from their gestures, thereby avoiding the necessity of input from users and eventually improving the communication mechanism with the robots. Difficulties in action recognition occur due to cluttered background, camera motion, occlusion, viewpoint changes and geometric and photometric variances of objects. Application areas for the human action recognition include virtual reality, games, video indexing, teleconferencing, advanced user interfaces and video surveillance.

1.1. Motivation and proposed work

Much work has been done regarding action recognition (Aggarwal & Ryoo, 2011; Moeslund, Hilton, & Kruger, 2006; Poppe, 2010; Weinland, Ronfard, & Boyer, 2011). Some methods build a model first and then recognize actions by tracking the changes of different limbs in either 2D (Ferrari, Marín-Jiménez, & Zisserman, 2009) or 3D (Horaud, Niskanen, Dewaele, & Boyer, 2009) spaces. These approaches are complex, due to the large variability of shapes and articulations of human bodies, fast motions, self occlusion, changes in appearance, etc. Other approaches employ gradient or intensity based features, e.g., the bag of words methods (Wei, Dacheng, & Yong, 2012) to recognize actions. The performance of the gradient

* Corresponding author.
E-mail addresses: atiqur.rahman@jcu.edu.au, shah0018@e.ntu.edu.sg (S.A. Rahman), insu.song@jcu.edu.au (I. Song), asmkleung@gmail.com (M.K.H. Leung), ickjai.lee@jcu.edu.au, shah0018@e.ntu.edu.sg (I. Lee), joanne.lee@jcu.edu.au (K. Lee).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.07.082

and intensity based methods depends on the detection of a sufficient number of stable interest points. However, detecting interest points could be difficult due to complex backgrounds. Silhouette-based methods are becoming popular due to their robustness to noise and because it is easier to extract regions of interest (ROI) than the space–time interest points (Wang & Suter, 2007a). Based on the dynamics of the actions, silhouette based methods can fall into two categories: *explicit* and *implicit*. In the explicit-model based methods, actions are treated as a composition of a sequence of poses (Wang & Suter, 2007a). First, each pose is described individually, and then the dynamics of poses are represented by, for example, Hidden Markov Model (HMM) and Dynamic Time Warping (DTW) to recognize actions. However, most of the explicit-model based techniques need to model each action individually, and usually explicit methods take a longer time in action classification than implicit methods. Implicit methods extract features from the combination of all the poses of an action sequence and generate action descriptors. The implicit modeling approaches have the advantage of simple and efficient action recognition, i.e., faster processing time, and can work with a small number of training data. However, most of the implicit methods require an additional step of dimensionality reduction by using different techniques, e.g., Principle Component Analysis (PCA) and Linear Discriminant Analysis (LDA)). Additionally, some of the systems require manual identification of cycle lengths (Kumar, Raman, & Sukavanam, 2011) or use sliding windows of fixed sizes (Gorelick, Blank, Shechtman, Irani, & Basri, 2007) to construct the action template.

In this paper, we propose an implicit-model based method that uses Negative Space Action Descriptors (NSAD) which form descriptions of actions using surrounding regions of poses. Since no individual features, such as hands and head, need to be identified and tracked, NSAD constructs much more efficiently than previous methods and is robust to noisy segmentation. In this paper, we show that the method can be applied to recognition of human actions, as well as activities of other subjects, e.g., fishes. The proposed system extracts features from the negative space, while other implicit methods extract features from the positive space. For action recognition, the silhouette of a subject is termed positive space, and the surrounding regions of the silhouette inside its bounding box are referred to as the negative space of the subject, as shown in Fig. 1. Basically, the negative space of a subject is the inverse set of the positive space (silhouette) of the subject. Now, the question is why negative space based action description is efficient? The answer is that negative space can describe poses by simple shapes which are naturally formed inside the bounding box, avoiding sophisticated and complex limb tracking or high dimensional polygonal approximation, which are used to describe poses in positive space based methods (Ikizler & Forsyth, 2008; Sigal & Black, 2006). For example, positive space based methods, such as feature tracking methods, first identify different limbs of the human body hierarchically by using some reference points

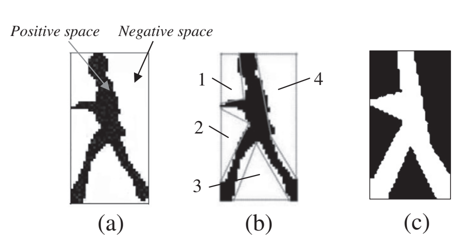

Fig. 1. (a) bounding box of a human silhouette. White regions inside the box are negative space, and the human silhouette is the positive space; (b) negative space regions can be viewed as triangles or quadrangles; (c) poses are represented *y* using triangles or quadrangles only.

and then track the limbs to recognize actions, which is a complex process. On the other hand, the negative space based methods only need to identify the bounding box of the human silhouette, which is much simpler than identifying and tracking different limbs, and the negative space regions formed inside the bounding box can be described using simple shapes. In Fig. 1(b), each of the four negative space regions (1, 2, 3, 4) is represented by either triangles or quadrangles which are formed naturally by the bounding box and the silhouette. A pose can then be represented by using the simple shapes only, as shown in Fig. 1(c).

Another advantage of our system is that the method is relatively robust to partial occlusion and noisy segmentation, since negative space based pose description is less affected by partial occlusion and noisy segmentation (Rahman, Cho, & Leung, 2012a). Specially, in the case of noisy segmentation, negative space based methods are particularly better than other state of the art methods. Furthermore, since each negative space region can be described by simple shapes, e.g., triangles, quadrangles, our system describes poses by low dimensional and computationally efficient feature vectors. These low dimensional feature vectors are later used to form negative space action descriptors (NSADs) for describing actions. Hence, each NSAD describes an action with low dimensional features. This implies that NSAD does not require dimensionality reduction, whereas most of the other implicit methods require a significant amount of processing time for dimensionality reduction (Zhang, Zhao, & Xiong, 2010). Additionally, in our system, we propose a method based on signal processing technique to determine the cycle length of each action automatically, without any prior knowledge of the action.

The feature extraction method from poses proposed here is the same as our previous method (Rahman, Leung, & Cho, 2011), but the action representation technique proposed here is novel. The main contribution of this paper is the novel method of constructing action descriptors that generate features from negative space based pose descriptions which can efficiently describe poses using simple shapes anchored to the bounding box, whereas positive space based methods need to identify suitable reference points to perform complex limb tracking or high dimensional polygonal approximation. The new action descriptors can be used to recognize actions more efficiently with comparable accuracy than the state of the art methods. Furthermore, the proposed method has shown very good accuracy in the case of corrupted image sequences, especially for corrupted images due to noisy segmentation. Another contribution is to compute the number of frames in a cycle (cycle length) of an action automatically without any prior knowledge about the action. A shorter version of this work can be found in Rahman, Song, and Leung (2012b).

1.2. Paper organization

The remainder of this paper is organized as follows. In Section 2, previous work is reviewed. In Section, 3 we describe our system in detail, as to how it computes moving speeds, partitions negative space, extracts features, forms action descriptors, and classifies descriptors. Section 4 shows experimental results of our system and compares system performance with other methods. Section 5 concludes the paper.

2. Previous work

Many criteria could be used to classify previous work, for example, low level feature cues, type of models and dimensionality of the tracking space. We have chosen the first criterion, since we mainly focus on proposing simple and efficient features to describe actions. Existing work can be broadly divided into three major