Computer Vision and Image Understanding 115 (2011) 334–351

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Learning spatio-temporal dependency of local patches for complex motion segmentation

Jiang Xu [a,*], Junsong Yuan [b], Ying Wu [a]

[a] Department of EECS, Northwestern University, Evanston, IL 60208, USA
[b] School of EEE, Nanyang Technological University, Singapore 639798, Singapore

ARTICLE INFO

Article history:
Received 5 March 2010
Accepted 16 November 2010
Available online 9 December 2010

Keywords:
Motion segmentation
Learning
Motion profile symmetry correlation
Bipolar segmentation

ABSTRACT

Segmenting complex motion, such as articulated motion and deformable objects, can be difficult if the prior knowledge of the motion pattern is not available. We present a novel method for motion segmentation by learning the motion priors from exemplar motions to guide the segmentation. Instead of modeling the motion field explicitly, we decompose each video frame into a number of local patches and learn the spatio-temporal contextual relations among them, *e.g.*, if their motion relationships are consistent with that from the training data. Based on a novel motion feature to measure the relative motion of two patches, the SVM classifier learns their pairwise relationship. We convert the motion segmentation problem to a binary labeling problem, and propose an iterative solution to group the local patches whose motions are consistent. Compared with other approaches, such as the graph cut and normalized cut methods, this new method is computationally more efficient and is able to better handle the inaccurate inference of pairwise relationships. Results on both synthesized and real videos show that our method can learn to segment different types of complex motion patterns.

1. Introduction

Segmenting a scene based on its motion is an important problem in computer vision. A common approach to motion segmentation is to extract the local motion features (*e.g.*, optical flow [1]) of all the pixels, then perform grouping based on the coherence in these local motion features.

The coherence among these local motion features actually depends on the complexity in the motion of the object, and thus also on the semantic levels of the object. Conventionally, most motion segmentation methods are either based on the coherence in terms of simple motions (such as rigid or affine motion [2–4]), or based on the smoothness in the motion (*e.g.*, smooth motion layers, or Markov random fields [5–9]). Methods based on the coherence in simple motions may result in over-segmentation if the object exhibits more complex motion, *e.g.*, articulation or deformation.

However, these methods are confronted when we want to segment an object presenting more complex motion, *e.g.*, a shearing scissors or a walking person (Fig. 1). Although the motion of the parts can be simple, the entire motion of the object does have more degrees of freedom, and cannot be covered by simple rigid motions or smooth motions. This paper is concerned on a new learning-based solution to this problem.

This is a quite difficult problem as the motion can be quite complex. However, in reality, it is not uncommon that the complex motions may exhibit good structures, implying that the intrinsic dimension of these motions can be actually low. In the example of the shearing scissors, the opening–closing motion pattern is quite predictable so that a specific motion model can be easily specified. However, in the case of a walking person, although the cyclic motion pattern does have a low intrinsic dimension, it is very difficult, if not impossible, to have an explicit model for such a motion.

In this paper we propose to learn the complex motion from training examples for motion segmentation. The complex motion patterns are labeled in the training video. After learning, we can segment objects in video that exhibit the same motion pattern.

To achieve this goal, we decompose each video frame into a few local image patches and focus on learning the spatio-temporal relationship among the patches. Specifically, suppose i and j are two local patches, we formulate the pairwise relationship $f_{i,j}$ as a binary classification problem: $f_{i,j} = 1$ if two patches belong to the target motion, and $f_{i,j} = -1$ otherwise. Given a test video where the pairwise relations among the patches are inferred, a novel segmentation approach is further proposed in this paper to find the most stable group iteratively. As the pairwise relationships are mostly consistent to each other in the stable group, it provides a robust segmentation of the target motion pattern, despite the noisy and cluttered backgrounds.

* Corresponding author.
E-mail address: jxu323@eecs.northwestern.edu (J. Xu).

1077-3142/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2010.11.010

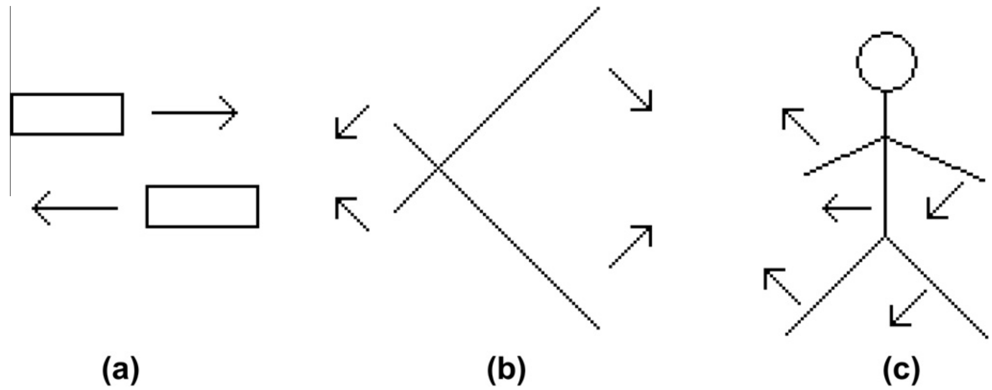

Fig. 1. Examples of motion that cannot be segmented by traditional methods. (a) Center-symmetrical motion, such as two trains passing by. (b) Mirror-symmetrical motion, such as a shearing scissors. (c) Articulated motion, such as human walking. The intrinsic dimensionality of the motion patterns above is not very high. However, we cannot simply use the rigid-body prior or smooth prior to segment those structured motions out.

To provide robust motion segmentation, we address the following two critical questions:

(1) How can we learn the pairwise relationship $f_{i,j}$ between two patches i and j?

Given two local patches, i and j, we need to identify whether their relative motion is consistent with that from the training examples. This is a two-class classification problem, as $f_{i,j}$ is either 1 or -1. To learn the pairwise relation $f_{i,j}$, a novel spatio-temporal contextual feature, called *motion profile symmetry correlation*, is presented to characterize the relative motion of the two patches i and j. This type of feature can be used to describe complex motion, such as opposite motion and mirror-symmetric motion. In the training sequences, we manually label the region of the target motion pattern, in order to obtain the ground truth of the pairwise relations $f_{i,j}$. Then a support vector machine (SVM) is applied to learn the binary relation $f_{i,j}$ using the proposed feature.

(2) How to segment the target motion pattern given the pairwise relation $f_{i,j}$?

To segment a frame in a test sequence, we use all of the local patches x_i to construct a graph, where each x_i corresponds to a

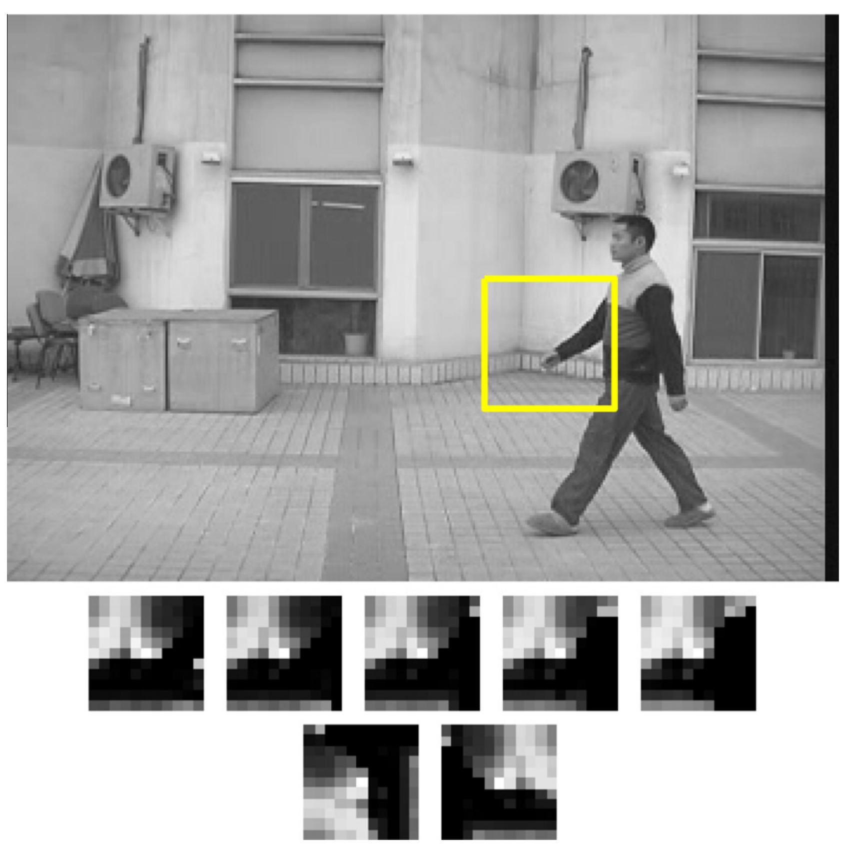

Fig. 2. Top row: the region for motion profile extraction. Middle row: the motion profiles of the patch at the center of the region, from frame $t-2$ to frame $t+2$. From the motion profiles, we can observe the arm region is moving up-left in those frames. Bottom row: the steering and reflection motion profiles. The left one is ${}^{90}P_i^t$, and the right one is $\left({}^{90}P_i^t\right)^T$.