Computer Vision and Image Understanding 117 (2013) 587–602

Contents lists available at SciVerse ScienceDirect

Computer Vision and Image Understanding

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

View invariant action recognition using weighted *fundamental ratios* ☆

Nazim Ashraf [a,*], Yuping Shen [a,b], Xiaochun Cao [a,c], Hassan Foroosh [a,d,1]

[a] College of Engineering and Computer Science, Computational Imaging Lab, University of Central Florida, 4000 Central Florida Blvd., Orlando, FL 32816, USA
[b] Advanced Micro Devices, Quadrangle Blvd., Orlando, FL 32817, USA
[c] State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
[d] Department of EECS, Computational Imaging Laboratory, University of Central Florida, Orlando, FL 32816, USA

ARTICLE INFO

Article history:
Received 10 February 2012
Accepted 19 January 2013
Available online 5 February 2013

Keywords:
View invariance
Pose transition
Action recognition
Action alignment
Fundamental ratios

ABSTRACT

In this paper, we fully investigate the concept of *fundamental ratios*, demonstrate their application and significance in view-invariant action recognition, and explore the importance of different body parts in action recognition. A moving plane observed by a fixed camera induces a fundamental matrix **F** between two frames, where the ratios among the elements in the upper left 2×2 submatrix are herein referred to as the *fundamental ratios*. We show that *fundamental ratios* are invariant to camera internal parameters and orientation, and hence can be used to identify similar motions of line segments from varying viewpoints. By representing the human body as a set of points, we decompose a body posture into a set of line segments. The similarity between two actions is therefore measured by the motion of line segments and hence by their associated *fundamental ratios*. We further investigate to what extent a body part plays a role in recognition of different actions and propose a generic method of assigning weights to different body points. Experiments are performed on three categories of data: the controlled CMU MoCap dataset, the partially controlled IXMAS data, and the more challenging uncontrolled UCF-CIL dataset collected on the internet. Extensive experiments are reported on testing (i) view-invariance, (ii) robustness to noisy localization of body points, (iii) effect of assigning different weights to different body points, (iv) effect of partial occlusion on recognition accuracy, and (v) determining how soon our method recognizes an action correctly from the starting point of the query video.

1. Introduction

The perception and understanding of human motion and action is an important area of research in computer vision that plays a crucial role in various applications such as surveillance, human–computer interaction (HCI), ergonomics, etc. In this paper, we focus on the recognition of actions in the case of varying viewpoints and different and unknown camera intrinsic parameters. The challenges to be addressed in action recognition include perspective distortions, differences in viewpoints, anthropometric variations, and the large degrees of freedom of articulated bodies [1]. The literature in human action recognition has been extremely active in the past two decades and significant progress has been made in this area [2–5]. Action can be regarded as a collection of 4D space–time data observed by a perspective video camera. Due to image projection, the 3D Euclidean information is lost and projectively distorted, which makes action recognition rather challenging, especially for varying viewpoints and different camera parameters. Another source of challenge is the irregularities of human actions due to a variety of factors such as age, gender, circumstances, etc. The timeline of action is another important issue in action recognition. The execution rates of the same action in different videos may vary for different actors or due to different camera frame rates. Therefore, the mapping between same actions in different videos is usually highly non-linear.

To tackle these issues, often simplifying assumptions are made by researchers on one or more of the following aspects: (1) camera model, such as scaled orthographic [6] or calibrated perspective camera [7]; (2) camera pose, i.e. little or no viewpoint variations; (3) anatomy, such as isometry [8], coplanarity of a subset of body points [8], etc. Human action recognition methods start by assuming a model of the human body, e.g. silhouette, body points, stick model, etc., and build algorithms that use the adopted model to recognize body pose and its motion over time. Space–time features are essentially the primitives that are used for recognizing actions, e.g. photometric features such as the optical flow [9–11] and the local space–time features [12,13]. These photometric features can be affected by luminance variations due to, for instance, camera zoom or pose changes, and often work better when the motion is small or incremental. On the other hand, salient geometric features such as silhouettes [14–18] and point sets [8,19] are less sensitive to photometric variations, but require reliable tracking. Silhouettes

☆ This paper has been recommended for acceptance by J.K. Aggarwal.

* Corresponding author.
E-mail address: nazim@cs.ucf.edu (N. Ashraf).
[1] Assistant Professor, Department of Computer Science, FC College (A Chartered University).

1077-3142/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.cviu.2013.01.006

are usually stacked in time as 2D [16] or 3D object [14,18], while point sets are tracked in time to form space–time curves. Some existing approaches are also more holistic and rely on machine learning techniques, e.g. HMM [20], SVM [12], etc. As in most exemplar-based methods, they rely on the completeness of the training data, and to achieve view-invariance, are usually expensive, as it would be required to learn a model from a large dataset.

1.1. Previous work on view-invariance

Most action recognition methods adopt simplified camera models and assume fixed viewpoint or simply ignore the effect of viewpoint changes. However, in practical applications such as HCI and surveillance, actions may be viewed from different angles by different perspective cameras. Therefore, a reliable action recognition system has to be invariant to the camera parameters or viewpoint changes. View-invariance is, thus, of great importance in action recognition, and has started receiving more attention in recent literature.

One approach to tackle view-invariant action recognition has been based on using multiple cameras [20–22,7]. Campbell et al. [23] use stereo images to recover a 3D Euclidean model of the human subject, and extract view invariance for 3D gesture recognition; Weinland et al. [7] use multiple calibrated and background-subtracted cameras, and they obtain a visual hull for each pose from multi-view silhouettes, and stack them as a motion history volume, based on which Fourier descriptors are computed to represent actions. Ahmad et al. [20] build HMMs on optical flow and human body shape features from multiple views, and feed a test video sequence to all learned HMMs. These methods require the setup of multiple cameras, which is quite expensive and restricted in many situations such as online video broadcast, HCI, or monocular surveillance.

A second line of research is based on a single camera and is motivated by the idea of exploiting the invariants associated with a given camera model, e.g. affine, or projective. For instance, Rao et al. [24] assume an affine camera model, and use dynamic instants, i.e. the maxima in the space–time curvature of the hand trajectory, to characterize hand actions. The limit with this representation is that dynamic instants may not always exist or may not be always preserved from 3D to 2D due to perspective effects. Moreover the affine camera model is restrictive in most practical scenarios. A more recent work reported by Parameswaran et al. [8] relaxes the restrictions on the camera model. They propose a quasi-view-invariant 2D approach for human action representation and recognition, which relies on the number of invariants in a given configuration of body points. Thus a set of projective invariants are extracted from the frames and used as action representation. However, in order to make the problem tractable under variable dynamics of actions they introduced heuristics, and made simplifying assumptions such as isometry of human body parts. Moreover, they require that at least five body points form a 3D plane or the limbs trace planar area during the course of an action. [25] described a method to improve discrimination by inferring and then using latent discriminative aspect parameters. Another interesting approach to tackle unknown views has been suggested by [26], who use virtual views, connecting the action descriptors extracted from source view to those extracted from target view. Another interesting approach is [27], who used a bag of visual-words to model an action and present promising results.

Another promising approach is based on exploiting the multi-view geometry. Two subjects in the same exact body posture viewed by two different cameras at different viewing angles can be regarded as related by the epipolar geometry. Therefore, corresponding poses in two videos of actions are constrained by the associated fundamental matrices, providing thus a way to match poses and actions in different views. The use of fundamental matrix in view invariant action recognition is first reported by Syeda-Mahmood et al. [28] and later by Yilmaz et al. [18,19]. They stack silhouettes of input videos into space–time objects, and extract features in different ways, which are then used to compute a matching score based on the fundamental matrices. A similar work is also presented in [29], which is based on body points instead of silhouettes. A recent method [30] uses probabilistic 3D exemplar model that can generate 2D view observations for recognition.

1.2. Our approach

This work is an extension of [31], which introduced the concept of *fundamental ratios* that are invariant to rigid transformations of camera, and were applied to action recognition. We make the following main extensions: (i) Instead of looking at fundamental ratios induced by triplets of points, we look at fundamental ratios induced by line segments. This, as we will later see, introduces more redundancy and results in better accuracy. (ii) It has been long argued in the applied perception community [32] that humans focus only on the most significant aspects of an event or action for recognition, and do not give equal importance to every observed data point. We propose a new generic method of learning how to assign different weights to different body points in order to improve the recognition accuracy by using a similar focusing strategy as humans; (iii) We study how this focusing strategy can be used in practice when there is partial but significant occlusion; (iv) We investigate how soon after the query video starts our method is capable of recognizing the action - an important issue never investigated by others in the literature; and (v) our experiments in this paper are more extensive than [31] and include larger set of data with various levels of difficulty.

The rest of the paper is organized as follows: In Section 2, we introduce the concept of *fundamental ratios*, which are invariant to rigid transformations of camera, and describe how they may be used for action recognition in Section 3. Then in Section 4, we focus on how we can weigh different body parts for better recognition. We present our extensive experimental evaluation in Section 5, followed by discussions and conclusion in Section 6.

2. Fundamental ratios

In this section, we establish specific relations between the epipolar geometry induced by line segments. We derive a set of feature ratios that are invariant to camera intrinsic parameters for a natural perspective camera model of zero skew and unit aspect ratio. We then show that these feature ratios are projectively invariant to similarity transformations of the line segment in the 3D space, or equivalently invariant to rigid transformations of camera.

Proposition 1. *Given two cameras* $\mathbf{P}_i \sim \mathbf{K}_i[\mathbf{R}_i|\mathbf{t}_i]$, $\mathbf{P}_j \sim \mathbf{K}_j[\mathbf{R}_j|\mathbf{t}_j]$ *with zero skew and unit aspect ratio, denote the relative translation and rotation from* $\mathbf{P}_i$ *to* $\mathbf{P}_j$ *as* $\mathbf{t}$ *and* $\mathbf{R}$ *respectively, then the upper* 2×2 *submatrix of the fundamental matrix between two views is of the form*

$$\mathbf{F}^{2\times 2} \sim \begin{bmatrix} \epsilon_{1st}\mathbf{t}^s\mathbf{r}_1^t & \epsilon_{1st}\mathbf{t}^s\mathbf{r}_2^t \\ \epsilon_{2st}\mathbf{t}^s\mathbf{r}_1^t & \epsilon_{2st}\mathbf{t}^s\mathbf{r}_2^t \end{bmatrix}, \tag{1}$$

where $\mathbf{r}_k$ *is the kth column of* $\mathbf{R}$, *the superscripts* $\mathbf{s}, \mathbf{t} = 1, \ldots, 3$ *indicate the element in the vector, and* ϵ_{rst}, $\mathbf{r} = 1, 2$ *is a permutation tensor.* [1]

Remark 1. The ratios among elements of $\mathbf{F}^{2\times 2}$ are invariant to camera calibration matrices $\mathbf{K}_i$ and $\mathbf{K}_j$.

[1] The use of tensor notation is explained in details in [33, p. 563].