Pattern Recognition Letters 51 (2015) 8–15

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Action recognition in still images by learning spatial interest regions from videos☆

Abdalrahman Eweiwi[a,*], Muhammad Shahzad Cheema[a], Christian Bauckhage[a,b]

[a] Bonn Aachen International Center for IT, University of Bonn, Germany
[b] Fraunhofer Institute for Intelligent Analysis and Information Systems, Sankt Augustin, Germany

ARTICLE INFO

Article history:
Received 18 September 2013
Available online 14 August 2014

Keywords:
Action recognition
Non-negative matrix factorization
Pose estimation
Optical flow

ABSTRACT

A common approach to human action recognition from still images consists in computing local descriptors for classification. Typically, these descriptors are computed in the vicinity of key points which either result from running a key point detector or from dense sampling of pixel coordinates. Such key points are not a priory related to human activities and thus might not be very informative with regard to action recognition. Several recent approaches, on the other hand, are based on learning person–object interactions and saliency maps in images. In this article, we investigate the possibility and applicability of identifying action-specific points or regions of interest in still images based on information extracted from video data. In particular, we propose a novel method for extracting spatial interest regions where we apply non-negative matrix factorization to optical flow fields extracted from videos. The resulting basis flows are found to indicate image regions that are specific to certain actions and therefore allow for an informed sampling of key points for feature extraction. We thus present a generative model for action recognition in still images that allows for characterizing joint distributions of regions of interest, local image features (visual words), and human actions. Experimental evaluation shows that (a) our approach is able to extract interest regions that are highly correlated to those body parts most relevant for different actions and (b) our generative model achieves high accuracy in action classification.

1. Introduction

Throughout the last decade, the problem recognizing human activities from still images has found considerable attention. Corresponding research is motivated by promising applications in areas such as automatic indexing of very large image repositories but is also expected to contribute to the solution of problems in automatic scene description, context dependent object recognition, or pose estimation [33,19,38].

Based on the underlying problem formulation, approaches to action recognition can be categorized into two main classes: (a) pose-based and (b) bag-of-features (BoF) approaches. Motivated by the idea of *poselets* [4], a notion of distributed part-based templates, pose-based approaches have recently been met with rekindled interest [40,27,43]. But the construction of poselets still requires a cumbersome procedure of manual annotation which impedes their use on large training sets. BoF approaches based on local descriptors are known for their state-of-the-art performance in object recognition and therefore have been adapted to action recognition [9]. However, local image descriptors are typically computed in the vicinity of key points that result from low-level signal analysis or from dense or random sampling and are therefore uninformative and independent of the activity depicted in an image. Several recent approaches are based on the idea of learning interaction between people and objects using saliency maps in images [31] or videos [6,5]. In this article, we presents a novel approach in this direction.

Most physical activities of people are characterized by articulation and movement of different body parts. And although activities are inherently dynamic, most people can easily infer human activities in still images just by looking at posture or configuration of particular body parts. Consider, for instance, the images shown in Fig. 1 which we can interpret even without having a full view of the human body. This raises the question if it is possible to automatically learn or identify action-specific, informative, regions of interest in still images without having to rely on exhaustive mining of low-level image descriptors or labor-intensive annotations?

In an attempt to answer this question, we propose an efficient yet effective approach towards automatic learning of action

☆ This paper has been recommended for acceptance by S. Wang.

* Corresponding author. Tel.: +49 228 2699131.
E-mail addresses: eweiwi@bit.uni-bonn.de (A. Eweiwi), cheema@bit.uni-bonn.de (Muhammad Shahzad Cheema), christian.bauckhage@iais.fraunhofer.de (Christian Bauckhage).

http://dx.doi.org/10.1016/j.patrec.2014.07.017
0167-8655/

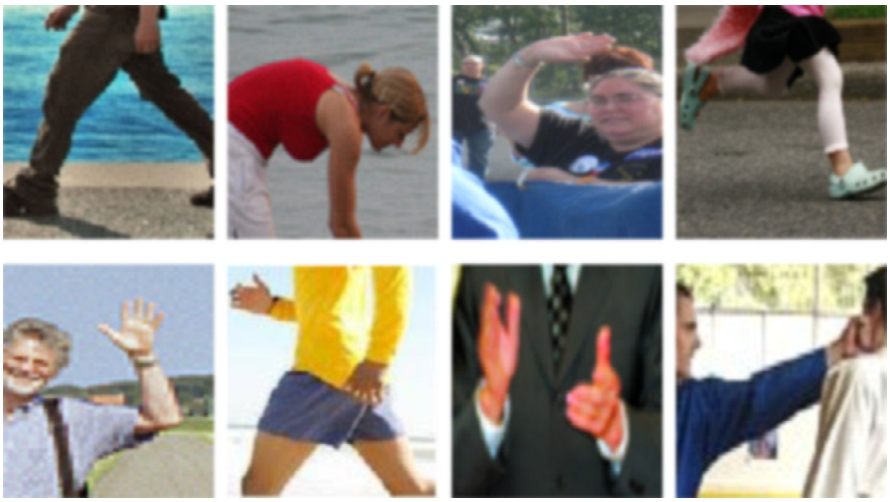

Fig. 1. Examples of image patches in which we can recognize human activities even though a view of the whole body is not available.

specific regions of interest in still images. Based on the observation that activities are temporal phenomena, we make use of information that is available from video analysis. Fig. 2, shows a diagram of the components of our approach towards determining action-specific regions of interest regions and subsequent image classification.

Given videos that show human activities, we compute optical flow fields and consider the magnitudes of flow vectors in each frame of a video. Given a collection of frames of flow magnitudes, we then apply non-negative matrix factorization (NMF) and obtain basis flows. These basis flows are indicative of the position and configuration of different limbs or body part whose motion characterizes certain activities. Viewed as images, the basis flows indicate action specific regions of interest and therefore allow for an informed sampling of interest points or regions for subsequent feature extraction. For action classification in still images, we devise a generative probabilistic model that characterizes joint distributions of regions of interest, local image features (visual words) and human actions.

To evaluate the usefulness of regions of interest contained in basis flows, we consider correspondences between regions of interest that were automatically learned from videos and manually annotated locations of human body parts that are available from an independent set of still images. Our empirical results reveal a high correlation between extracted interest regions and those body parts that are most relevant for different actions. Below, we show that, even in the absence of any annotation of joints or body parts, our generative model achieves a high accuracy in action classification.

The major contributions of this article are the following: (i) we propose a novel scheme for extracting discriminative spatial regions for action recognition in still images using simple videos; (ii) we apply NMF to determine action specific regions of interest from motion flows; (iii) we incorporate action saliency maps based on videos and local spatial features of action images in a Bayesian framework for human action classification.

Our presentation proceeds as follows. In Section 2, we review related work on human action recognition. Section 3 describes our method of learning action specific interest regions from videos. In Section 4, we present a generative model for action classification. Section 5 provides an experimental evaluation with respect to the location of human body parts. Finally, Section 6 summarizes our work and results.

2. Related work

As an exhaustive review of work on visual activity recognition is beyond the scope of this article, we restrict our discussion to the

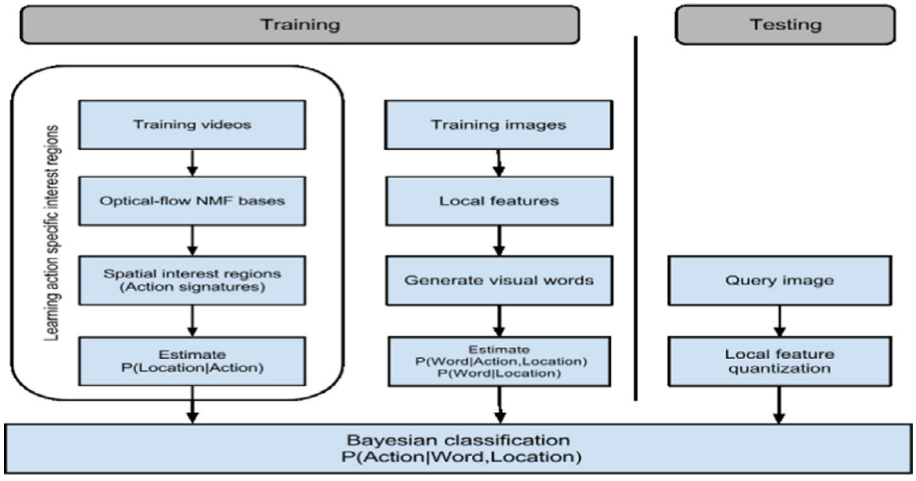

Fig. 2. General diagram of our approach. In the training phase, we learn the actions' priors P(Location | Action) from training videos, and codebook prior P(Word | Action, Location) from training images in order to perform human action recognition for new test queries in a fully Bayesian framework.