Journal of Visual Languages and Computing 28 (2015) 163–175

Contents lists available at ScienceDirect

Journal of Visual Languages and Computing

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

Human action recognition in still images using action poselets and a two-layer classification model ☆

ByoungChul Ko *, JuneHyeok Hong, Jae-Yeal Nam

Department of Computer Engineering, Keimyung University, Sindang-dong, Dalseo-gu, Daegu 704-701, Republic of Korea

ARTICLE INFO

Article history:
Received 22 August 2014
Received in revised form
14 January 2015
Accepted 15 January 2015
Available online 30 January 2015

Keywords:
Action recognition
Two-layer classification model
Spatial poselet activation vector
OCS-LBP
Random forest

ABSTRACT

Human action recognition in still images provides useful information for use in a wide range of computer vision applications. Because motion information cannot be estimated from a single image, action recognition in still images remains a challenging problem.

In this paper, we propose poselet based action recognition methods to infer human action using a two-layer classification model. First, poselets are built using the annotated data of joint locations of people and the proposed Hausdorff distance. Each poselet, consisting of a feature vector, is trained using the first-layer classifier based on random forest classification to find the proper location. From trained poselet detectors, we construct spatial poselet activation vectors (SPAVs) using the voting scores of poselets. A second-layer classifier, which takes aggregating SPAVs of the first-layer classifiers as input, trains a final multi-class classifier. During the testing phase, the input window, which includes a human region, is applied to the first-layer classifier; the aggregating output of the first-layer is applied to the second-layer classifier. After calculating scores for all the *c*-action classes, the final action class is selected as the one that has the maximum score. Experimental results showed that the recognition performance and processing times of the proposed method was better than those of previous methods.

1. Introduction

While latest high-resolution digital cameras and smart phones are able to capture high-quality digital images, and have enabled users to capture pictures anywhere and at any time, automatic understanding of images is still in its initial steps because certain high-level vision problems, such as scene analysis, object detection, human detection, pose estimation, and action recognition, still need to be resolved. Human action recognition is one of the most important issues in computer vision owing to its potential to provide useful meta-data to a wide range of applications such as video surveillance, gesture recognition, and human–computer interaction.

Most traditional methods of recognizing human actions use motion information in videos [1–3] because it often provides discriminative cues for action classification. In general, however, motion information cannot be estimated from only a single image. Several types of actions, such as 'standing on the road,' 'reading a book,' 'phoning,' and 'taking a photograph' are static in nature as shown in Fig. 1. The static pose and appearance of a person are valuable cues for inferring the action even if video is available.

The following problems need to be addressed before reliable action recognition using only a single image can be achieved [4]:

- The lack of region model in a single image precludes discrimination of foreground and background objects.

☆ This paper has been recommended for acceptance by Shi Kho Chang.

* Corresponding author. Tel.: +82 53 580 5235; fax: +82 53 580 6275.
E-mail addresses: niceko@kmu.ac.kr (B. Ko), pulsemate@kmu.ac.kr (J. Hong), jynam@kmu.ac.kr (J.-Y. Nam).

http://dx.doi.org/10.1016/j.jvlc.2015.01.003
1045-926X/© 2015 Elsevier Ltd. All rights reserved.

- The presence of articulation makes the recognition problem much harder.

1.1. Related works

Action recognition in still images is a very challenging problem. Interest in this problem has increased over the last few years with several recent studies [4–21] attempting to overcome the limitations of still images so that dependable action recognition results can be achieved. Methods used to recognize human actions in still images are classified into four main categories according to body pose as presented in Table 1.

The first category of methods of human pose estimation includes model-based approaches [5–6]. This method predefines an explicitly known parametric body model and solves the pose recovery problem by matching the pose variables to a forward rendered human model based on labeled extracted features. Gupta et al. [6] presented a graphical model for modeling human-object interactions. The nodes in the model represent the perceptual analyses corresponding to the recognition of objects, reach motions, manipulation motions, and object reactions. They used an edge and silhouette based likelihood representation for body parts and implemented a variant of pictorial structures for estimating the 2D pose of the upper body. The main drawbacks of the model based approach include initialization of the parametric body model and its pose.

The second category is example-based methods [7–9], which store a collection of images or image features with their corresponding pose description. These methods do not require a global modeling structure of the initialization/parameter relationship and use classic learning methods such as the *k*-nearest neighbor rule and locally-weighted regression due to their simplicity [7]. Shakhnarovich et al. [7] introduced a pose estimation method by using hashing-based search techniques to rapidly find relevant pose examples in a large database of image data and estimated the parameters for the input using a local model learned from those pose examples. Poppe et al. [8] compared three shape descriptors – Fourier descriptors, shape contexts, and Hu moments – that were used in the encoding of human silhouettes. An example based approach was taken to recover upper body poses from these descriptors. Poppe et al. performed experiments with deformed silhouettes to test each descriptor's robustness against variations in body dimension, viewpoint, and noise. Wang et al. [9] presented a method for discovering classes of actions in collections of still images by clustering images of people in similar body poses. Because computing the distance between a pair of images using a linear programming relaxation technique is a computationally expensive process, they employed a fast pruning method based on shape contexts to speed up the search for similar images. The major drawbacks of example-based approaches are their computational complexity of similarity searches in high-dimensional spaces and their propensity to include very large data sets.

The third category is pictorial-structure-based methods [4,10–15]. These methods mainly rely on pose representations as a cue for action recognition by finding body parts and constructing the overall pose using prior knowledge of

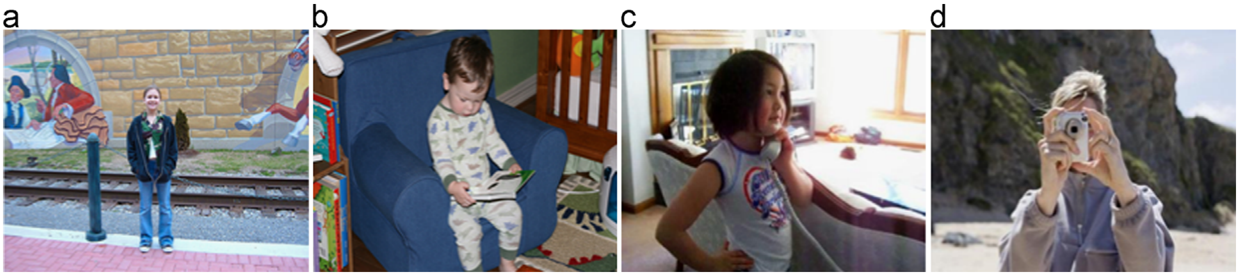

Fig. 1. Examples of static actions associated with specific poses: (a) standing on the road, (b) reading a book, (c) phoning, and (d) taking a photograph.

Table 1
The representative research categories of action recognition in still images.

Research categories	Characteristics
Model-based approaches [5,6]	• Predefining an explicitly known parametric body model • Matching the pose variables to a forward rendered human model
Example-based approaches [7–9]	• Storing a collection of images or image features with their corresponding pose description • Using classic learning methods such as the *k*-nearest neighbor rule and locally-weighted regression
Pictorial-structure-based approaches [4,10–15]	• Relying on pose representations as a cue for action recognition by finding body parts • Constructing the overall pose using prior knowledge of the human body structure
Poselet-based approaches [16–19]	• Representing poselets as a configuration of body part locations • Training each poselet using annotated 3D images