Chapter 91
Recognizing Human Actions in Basketball Video Sequences on the Basis of Global and Local Pairwise Representation

Masaki Takahashi
NHK Science and Technical Research Laboratories, Japan

Masahide Naemura
NHK Science and Technical Research Laboratories, Japan

Mahito Fujii
NHK Science and Technical Research Laboratories, Japan

James J. Little
The University of British Columbia, Canada

ABSTRACT

A feature-representation method for recognizing actions in sports videos on the basis of the relationship between human actions and camera motions is proposed. The method involves the following steps: First, keypoint trajectories are extracted as motion features in spatio-temporal sub-regions called "spatio-temporal multiscale bags" (STMBs). Global representations and local representations from one sub-region in the STMBs are then combined to create a "glocal pairwise representation" (GPR). The GPR considers the co-occurrence of camera motions and human actions. Finally, two-stage SVM classifiers are trained with STMB-based GPRs, and specified human actions in video sequences are identified. An experimental evaluation of the recognition accuracy of the proposed method (by using the public OSUPEL basketball video dataset and broadcast videos) demonstrated that the method can robustly detect specific human actions in both public and broadcast basketball video sequences.

DOI: 10.4018/978-1-5225-5204-8.ch091

1. INTRODUCTION

Recognition of human actions in video sequences (hereafter, "human-action recognition") is in great demand in fields such as video retrieval and sports-video analysis (Cedras & Shah, 1995; Ke, Thuc, Lee, Hwang, Yoo, & Choi, 2013). However, problems such as appearance changes, occlusions, motion blur, and camera motions make action recognition very challenging.

Particularly in the case of sports-video analysis, camera motion should be considered because it causes noise features in background regions. However, especially in broadcast sports videos shot by professional camera operators, some camera motions are triggered by human actions. For example, during golf programs, the camera is usually zoomed on the golf ball just after a putt has been taken. During basketball programs, the camera tends to be panned just after a shot has been taken because the offence tends to shift to the defending team. A combination of local features obtained from human actions and global features obtained from camera motions would thus be useful for sports-video analysis.

Image-recognition technology has been dramatically improved by using a combination of local-feature descriptors, such as SIFT, and feature-representation techniques, such as the bag-of-features (BoF) approach (Csurka, Dance, Fan, Willamowski, & Bray, 2004). This combined framework can also be applied to action recognition, and many studies have tried to identify specific human actions on the basis of the framework (Poppe, 2010; Aggarwal & Ryoo, 2011; Hassner, 2013).

However, to compensate for its scale-invariant characteristic, the normal BoF approach neglects positional information of local features, which is especially useful for analyzing broadcast videos (which tend to be shot in typical compositions). The spatial pyramid, which embeds the positional information of local features into feature representation, is one solution to consider positional information of features (Lazebnik, Schmid, & Ponce, 2006). Besides positional information, temporal features should be considered to handle objects' motion in video sequences (Matikainen, Hebert, & Sukthankar, 2010). Some methods that use a spatio-temporal pyramid for action recognition have been reported (Wang, Chen, & Wu, 2011; Choi, Jeon, & Lee, 2008).

Feature representation based on the BoF approach has been further improved (Farquhar, Szedmak, Meng, & Shawe-Taylor, 2005). For example, the Fisher kernel paradigm, which uses Gaussian mixture model (GMM), is a promising approach for representing local descriptors (Perronnin & Dance, 2007; Perronnin, Sánchez, & Mensink, 2010). The vector of locally aggregated descriptors (VLAD), which is a simple version of the Fisher kernel framework, is another promising approach due to its rich amount of information in spite of its low number of dimensions (J´egou, Douze, Schmid, & P´erez, 2010).

In this study, considering both global and local motion features, a new feature-representation method, which is suitable for human-action recognition in broadcast sports videos, is proposed. With this method, keypoint trajectory features are extracted from spatio-temporal sub-regions called "spatio-temporal multiscale bags" (STMBs) and represented by using VLAD. The authors believe this is the first attempt to apply a combination of a multiscale spatio-temporal strategy and a GMM-based feature-representation model to human-action recognition methods.

In addition, the "glocal pairwise representation" (GPR), which represents co-occurrence of local features both in the global area and in the local area, is proposed (glocal = global and local). Local features generated by camera operations were found to be effective in detecting specific human actions because some camera operations are triggered by a specific human action. A global-feature representation, which is created from the full spatial range, and a local-feature representation, which is created