Human Action Recognition in Video via Fused Optical Flow and Moment Features – Towards a Hierarchical Approach to Complex Scenario Recognition

Kathy Clawson¹, Min Jing², Bryan Scotney¹, Hui Wang², and Jun Liu²

¹ School of Computing and Information Engineering, University of Ulster
Cromore Road, Coleraine, BT52 1SA, UK
² School of Computing and Mathematics, University of Ulster
Shore Road, Newtownabbey, BT37 0QB, UK
{k.clawson,m.jing,bw.scotney,h.wang,j.liu}@ulster.ac.uk

Abstract. This paper explores using motion features for human action recognition in video, as the first step towards hierarchical complex event detection for surveillance and security. We compensate for the low resolution and noise, characteristic of many CCTV modalities, by generating optical flow feature descriptors which view motion vectors as a global representation of the scene as opposed to a set of pixel-wise measurements. Specifically, we combine existing optical flow features with a set of moment-based features which not only capture the orientation of motion within each video scene, but incorporate spatial information regarding the relative locations of directed optical flow magnitudes. Our evaluation, using a benchmark dataset, considers their diagnostic capability when recognizing human actions under varying feature set parameterizations and signal-to-noise ratios. The results show that human actions can be recognized with mean accuracy across all actions of 93.3%. Furthermore, we illustrate that precision degrades less in low signal-to-noise images when our moments-based features are utilized.

Keywords: Video pattern recognition, optical flow, Zernike moments, Hidden Markov Model.

1 Introduction

Scenario recognition and event detection in video streams has gained increasing attention in the computer vision research community due to the needs of many applications, such as surveillance for security [1]. The ubiquity of CCTV capture systems combined with the sheer volume of generated data has amplified the necessity for retrospective, automated (or semi-automated) evaluation, indexing, annotation and retrieval of surveillance video. Intelligent analysis of surveillance data faces unique challenges. Current CCTV systems suffer limitations, including the existence of: diversified and non-interoperable video archiving systems and proprietary technologies; massive volumes of recorded data; low resolution video; scene occlusion, background clutter; noise; low inter-class variance between events of interest; infrequently
occurring events of interest; limited training data; conflicting legal frameworks across geographically disparate locations; and privacy and ethical issues relating to semantic video analysis.

Our research constitutes part of the SAVASA project [1], which seeks to develop a standards–based approach to surveillance video archive, search and analysis. A key component of SAVASA is video analytics for automatic detection of concepts and scenarios, to allow video indexing via an operational ontology and enable authorised users to perform semantic queries (reducing their search space and retrieving potentially pertinent data). Within the literature, computer vision techniques can identify complex semantic events with generally low precision and recall. SAVASA proposes to improve upon the state-of-the-art by decomposing complex events into their atomic components and regarding high level event detection as a reasoning process [2] which searches for both the existence of multiple sub-events and for spatio-temporal sub-event associations / properties (Figure 1). The perceived benefits of a hierarchical approach include the ability to analyse data across multiple levels of detail, the ability to model sparsely occurring events based on their more frequently occurring compositions, the ability to allow for inaccurate / noisy lower level event detections via uncertainty reasoning [2], and the reduction in dependency on high volumes of training data.

This paper describes lower-level components of hierarchical scenario recognition, and constitutes video feature extraction and single layer recognition, with focus on human action recognition. The components presented, although not suitable for detection of complex events in isolation, will act as the lower level in a complete system for hierarchical scenario recognition as illustrated in Figure 1. In order to facilitate complex event detection, it is desirable to utilize features which are robust given a variety of image conditions, including noise. The main contribution of this work is an evaluation of the applicability of motion features, specifically optical flow features, as the sole basis of feature generation. We propose a set of fused features which combine and build upon existing features in the literature, and consider the robustness of our methodology across varying signal-to-noise (SNR) ratios. We wish to reduce a known limitation of classification using optical flow features, specifically that degradation occurs in reliability of pixel-wise motion vectors given low SNR ratios or given motion patterns which defy underlying assumptions upon which flow has been calculated (for example that subsequent frames are in approximate registration). For this reason, we follow the methodology of Efros et al [3] and regard the complete set of motion vectors as a spatial pattern of noisy measurements which have been smoothed and aggregated, as opposed to (more commonly) viewing them as precise pixel displacements. We extend upon this idea, and generate moment-based features which are subsequently used for classification.

The remainder of this paper is structured as follows. We present a brief overview of related literature in Section 2. Our methodology for atomic classification is described in Section 3. Our experimental overview and results are given in Section 4, and conclusions are offered in Section 5.