Chapter 9

Fast Categorisation of Articulated Human Motion

Konrad Schindler
TU Darmstadt, Germany

Luc van Gool
ETH Zürich, Switzerland ESAT/PSI-IBBT, K. U. Leuven, Belgium

ABSTRACT

Visual categorisation of human motion in video clips has been an active field of research in recent years. However, most published methods either analyse an entire video and assign it a single category label, or use relatively large look-ahead to classify each frame. Contrary to these strategies, the human visual system proves that simple categories can be recognised almost instantaneously. Here we present a system for categorisation from very short sequences (“snippets”) of 1–10 frames, and systematically evaluate it on several data sets. It turns out that even local shape and optic flow for a single frame are enough to achieve ≈80-90% correct classification, and snippets of 5-7 frames (0.2-0.3 seconds of video) yield results on par with the ones state-of-the-art methods obtain on entire video sequences.

1 INTRODUCTION

Recognising human motion categories in monocular video is an important scene understanding capability, with applications in diverse fields such as surveillance, content-based video search, and human-computer interaction. By motion categories we mean a semantic interpretation of the articulated human motion. Most computer vision research in this context has concentrated on human action recognition, while we only see actions as one possible set of semantic categories, which can be inferred from the visual motion pattern. We will also show a more subtle example, in which emotional states are derived from body language.

Past research in this domain can be roughly classified into two approaches: one that extracts a global feature set from a video (Ali et al., 2007; Dollár et al., 2005; Laptev and Lindeberg, 2003; Wang and Suter, 2007), and using these features aims to assign a single label to the entire video (typically several seconds in length). This paradigm obviously requires that the observed motion category does not change during the duration of the video.
The other approach extracts a feature set locally for a frame (or a small set of frames), and assigns an individual label to each frame (Blank et al., 2005; Efros et al., 2003; Jhuang et al., 2007; Niebles and Fei-Fei, 2007). If required, a global label for the sequence is obtained by simple voting mechanisms. Usually these methods are not strictly causal: features are computed by analysing a temporal window centred at the current frame, therefore the classification lags behind the observation—to classify a frame, future information within the temporal window is required.

Both approaches have achieved remarkable results, but human recognition performance suggests that they might be using more information than necessary: we can correctly recognise motion patterns from very short sequences—often even from single frames.

1.1 Aim of This Work

The question we seek to answer is how many frames are required to categorise motion patterns? As far as we know, this is an unresolved issue, which has not yet been systematically investigated (in fact, there is a related discussion in the cognitive sciences, see section 4). However, its answer has wide-ranging implications. Therefore, our goal is to establish a baseline, how long we need to observe a basic motion, such as walking or jumping, before we can categorise it.

We will operate not on entire video sequences, but on very short sub-sequences, which we call snippets. In the extreme case a snippet can have length 1 frame, but we will also look at snippets of up to 10 frames. Note that in many cases a single frame is sufficient, as can be easily verified by looking at the images in Figure 1. The main message of our study is that very short snippets (1-7 frames), are sufficient to distinguish a small set of motion categories, with rapidly diminishing returns, as more frames are added.

This finding has important implications for practical scenarios, where decisions have to be taken online. Short snippets greatly alleviate the problem of temporal segmentation: if a person switches from one action to another, sequences containing the transition are potentially problematic, because they violate the assumption that a single label can be applied. When using short snippets, only few such sequences exist. Furthermore, short snippets enable fast processing and rapid attention switching, in order to deal with further subjects or additional visual tasks, before they become obsolete.

Figure 1. Examples from databases WEIZMANN (left), KTH (middle), and LPP A (right); Note that even a single frame is often sufficient to recognise an action, respectively emotional state.
Related Content

Conclusion and a Look to the Future
www.irma-international.org/chapter/conclusion-look-future/60052/

An Interactive Tool for Visualizing and Analyzing Spatio-Temporal Data
www.irma-international.org/chapter/an-interactive-tool-for-visualizing-and-analyzing-spatio-temporal-data/164626/

Optimization of Power Allocation in Multimedia Wireless Sensor Networks
www.irma-international.org/article/optimization-power-allocation-multimedia-wireless/78552/

From Image to XML: Monitoring a Page Layout Analysis Approach for the Visually Impaired
www.irma-international.org/article/from-image-to-xml/116731/

Using Grid Computing and Satellite Remote Sensing in Evapotranspiration Estimation
www.irma-international.org/chapter/using-grid-computing-and-satellite-remote-sensing-in-evapotranspiration-estimation/164636/