Semi-Supervised Multimodal Fusion Model for Social Event Detection on Web Image Collections

Zhenguo Yang, Department of Computer Science, City University of Hong Kong, Hong Kong
Qing Li, Department of Computer Science, City University of Hong Kong, Hong Kong
Zheng Lu, Department of Computer Science, City University of Hong Kong, Hong Kong
Yun Ma, Department of Computer Science, City University of Hong Kong, Hong Kong
Zhiguo Gong, Department of Computer and Information Science, University of Macao, Taipa, Macao
Haiwei Pan, College of Computer Science and Technology, Harbin Engineering University, Harbin, China
Yangbin Chen, Department of Computer Science, City University of Hong Kong, Hong Kong

ABSTRACT

In this work, the authors aim to detect social events from Web images by devising a semi-supervised multimodal fusion model, denoted as SMF. With a multimodal feature fusion layer and a feature reinforcement layer, SMF learns feature histograms to represent the images, fusing multiple heterogeneous features seamlessly and efficiently. Particularly, a self-tuning approach is proposed to tune the parameters in the process of feature reinforcement automatically. Furthermore, to deal with missing values in raw features, prior knowledge is utilized to estimate the missing ones as a preprocessing step, and SMF will further extend an extra attribute to indicate if the values in the fused feature are missing. Based on the fused expression achieved by SMF, a series of algorithms are designed by adopting clustering and classification strategies separately. Extensive experiments conducted on the MediaEval social event detection challenge reveal that SMF-based approaches outperform the baselines.

Keywords: Event Detection, Feature Representation, Multimedia Data Mining, Multimodal Fusion, Semi-supervised Learning, Social Media Analysis

DOI: 10.4018/IJMDEM.2015100101
INTRODUCTION

The rapid development of digital photo-capture devices and the popularity of social media sites, provide new ways in which people share their experiences. As a consequence, a large amount of user-contributed multimedia data collections have been generated on the Web. Taking Flickr (a public photo-sharing site) as an example, more than 1.8 million photos are uploaded to its site per day, on average. Such huge collections of images serve as a growing record of our experiences and environment, which are usually associated with a wide range of real-world concepts and activities, such as landmarks, points of interest (POIs), social events, etc. Particularly, social events (Petkos, Papadopoulos, & Kompatsiaris 2012) are events that are organized by people and attended mostly by people who are not directly involved in the organization of the events. Instances of such events could be concerts, soccer games, promotion talks, etc. Research on new methodologies for event detection from these image collections has bloomed splendidly, aiming to link the multimedia collections on the Web to various kinds of real-world events they depict.

Social event detection (SED) is part of a broader initiative called Topic Detection and Tracking (TDT) (Allan, 2002), and can be characterized as a clustering or classification task. Different from traditional TDT tasks that generally focus on text documents, SED tasks in user-contributed Web images are deemed to be more challenging due to the characteristic of possessing multiple feature modalities. Taking Flickr as an example, the data is usually associated with multiple heterogeneous features, such as time-stamp, user identity, location, textual content and visual content, making it difficult to be effectively exploited by traditional clustering or classification models. Generally, researchers usually adopt early fusion or late fusion strategies in order to bridge the semantic gap among the multimodal features. Late fusion (Snoek, Worring, & Smeulders, 2005) is expensive in terms of learning effort and the results may not be good since a single modality might be poor. Consequently, early fusion strategy is more widely used in scenarios such as cross media retrieval, multi-modal clustering, SED, etc.

Despite the popularity of the early fusion models, a major disadvantage they shared is they are usually computation-intensive. It limits them from being applied to the tasks of detecting events from social media sites, where the data are huge and rapidly updated. Furthermore, incomplete data is an unavoidable problem when dealing with real world datasets, which will affect the efficiency and accuracy of the learning processes. Taking Flickr as an example, merely about 20% of uploaded images possess location information, despite that location is effectiveness in measuring the similarities among the media data samples in the context of SED. Quite a few works discarded them from further analyses (Firan, Georgescu, Nejdl, & Paiu, 2010; Trad, Joly, & Boujemaa, 2011), which did not attempt to make full use of the multimodal features.

In this work, a semi-supervised multimodal fusion model (SMF) is devised for SED tasks. SMF consists of two layers, i.e., multimodal feature fusion layer aiming to fuse the multiple heterogeneous features, and feature reinforcement layer aiming to learn the similarity propagations among the multiple aspects. SMF exploits an image dictionary in the process of fusing multiple heterogeneous features, representing images as the histograms of their similarities to the dictionary patterns. A dictionary pattern in the context of SED tasks could be a number of representative images related to specific events. SMF exploiting feature histograms make it less sensitive to the raw features, which could be incomplete in reality. Such as the current SED dataset, 80% of location information and all the machine tags are unavailable. To deal with the missing data problem, prior knowledge is used for estimations, and SMF will further extend an extra attribute to indicate if the values in the fused feature are missing or not (Lo et al., 2009). In particular, a self-tuning approach is presented to make the parameters in the second layer tuned automatically, which is verified to be effective in real-world SED tasks. Based on the new