DMMs-Based Multiple Features Fusion for Human Action Recognition

Mohammad Farhad Bulbul, Department of Information Science, School of Mathematical Sciences and LMAM, Peking University, Beijing, China

Yunsheng Jiang, Department of Information Science, School of Mathematical Sciences and LMAM, Peking University, Beijing, China

Jinwen Ma, Department of Information Science, School of Mathematical Sciences and LMAM, Peking University, Beijing, China

ABSTRACT

The emerging cost-effective depth sensors have facilitated the action recognition task significantly. In this paper, the authors address the action recognition problem using depth video sequences combining three discriminative features. More specifically, the authors generate three Depth Motion Maps (DMMs) over the entire video sequence corresponding to the front, side, and top projection views. Contourlet-based Histogram of Oriented Gradients (CT-HOG), Local Binary Patterns (LBP), and Edge Oriented Histograms (EOH) are then computed from the DMMs. To merge these features, the authors consider decision-level fusion, where a soft decision-fusion rule, Logarithmic Opinion Pool (LOGP), is used to combine the classification outcomes from multiple classifiers each with an individual set of features. Experimental results on two datasets reveal that the fusion scheme achieves superior action recognition performance over the situations when using each feature individually.

Keywords: Action Recognition, Depth Motion Maps, Edge Oriented Histograms, Kernel-based Extreme Learning Machine, Local Binary Patterns

INTRODUCTION

Automatic human action/gesture recognition is an active research topic in the area of computer vision. Researchers are fueled by the increasing number of real-world applications including autonomous visual surveillance, video retrieval, human-computer interaction, health care, sports training, etc. (e.g., C. Chen, Liu, Jafari, & Kehtarnavaz, 2014a; C. Chen, Kehtarnavaz, & Jafari, 2014b). Human action recognition is very challenging due to the significant variations in human body sizes, appearances, postures, motions, clothing, camera motions, viewing angles, illumination changes, etc. Moreover, the complexity grows due to that the same action is performed differently by different persons, even for same person at different times.

DOI: 10.4018/IJMDEM.2015100102
A large portion of researchers have addressed this problem by using features extracted from 2D intensity images (Chaaraoui, Climent-Pérez, & Flórez-Revuelta, 2012; Poppe, 2010; Willem, Madasu, Boles, & Yarlagadda, 2010; H. Wang & Schmid, 2013). However, the 2D intensity images captured by the conventional RGB video cameras do not have enough information to perform the comprehensive analysis. Moreover, they are sensitive to lighting condition, and the process of identifying key points depends on the object texture instead of object geometry (L. Chen, Wei, & Ferryman, 2013). On the other hand, these intensity images have many obstacles to perform robust computer vision tasks such as background subtraction and object segmentation.

Recently, with the availability of low-cost depth cameras (e.g., Microsoft Kinect), some of the difficulties for intensity images have been alleviated. The outputs of depth cameras are called depth images (which are sometimes mentioned as depth maps or depth frames according to context). Depth images preserve the depth information corresponding to the distances from the surface of scene objects to the viewpoint (Shotton, et al., 2013). The pixels in a depth image indicate calibrated depths (i.e., depths in a scale) in the scene, instead of intensity or color. This depth information achieves an additional robustness to color information due to its invariant to illumination and textures changes (Zhu & Pun, 2013). Moreover, the depth data captures the 3D structure of the scene as well as the 3D motion of the subjects/objects in the scene. Therefore, depth cameras show many advantages over the conventional intensity cameras, such as working under low light conditions and even in darkness, estimating calibrated depth, being steady to color and texture variations, and giving solution of the silhouette problem in posture (Shotton, et al., 2013). They also remove many ambiguities in computer vision tasks like background subtraction and object segmentation.

This paper proposes an effective action recognition framework by fusing the outcomes of multiple classifiers, each of which has an individual features set. This type of fusion is essential, as often a single kind of features or feature-level fusion may not exhibit enough discriminatory power. Therefore, we combine the classification decisions from classifiers with three types of features that extracted from Depth Motion Maps (DMMs) (C. Chen, Liu, & Kehtarnavaz, 2013): i) Contourlet-based Histogram of Oriented Gradients (CT-HOG) (Farhad, Jiang, & Ma, 2015a), ii) Local Binary Patterns (LBP) (Ojala, Pietikäinen, & Mäenpää, 2002) and iii) Edge Oriented Histograms (EOH) (Conaire). More specifically, we first represent an action video sequence with three DMMs (see Section 3 for more details). Then, CT-HOG, LBP and EOH are computed on each DMM separately. Finally, three feature sets are fed into three Kernel-based Extreme Learning Machine (KELM) (Huang, Zhu, & Siew, 2006) classifiers to provide the probability outputs for each action. The obtained probability outputs are merged using Logarithmic Opinion Pool (LOGP) (Benediktsson & Sveinsson, 2003) and Majority Voting (MV) (Lam & Suen, 1997) decision rules to label the query sample. Overall, the decision-level fusion operates on probability outputs and fuses multiple decisions into a joint one.

The main contributions of this paper are summarized as follows:

1. We compute three feature descriptors employing DMMs, CT-HOG, LBP and EOH. DMMs are utilized to capture specific appearances and shapes in a depth video sequence. Then, CT-HOG, LBP and EOH are employed on DMMs to obtain contour, texture and edge features respectively. Here, all these features lead us to achieve a compact representation of DMMs and enhance discriminatory power for the recognition algorithm.

2. Decision-level fusion is employed to the extracted compact features. In the decision-level fusion, more than one decision strategies are utilized to merge the probability outputs of each classification.
IPTV Challenges and Solutions in Metro Networks
Sajjad Zare and Akbar Ghaffarpour Rahbar (2012). *Advancements in Distributed Computing and Internet Technologies: Trends and Issues*  (pp. 40-63).
[www.igi-global.com/chapter/iptv-challenges-solutions-metro-networks/59677?camid=4v1a](www.igi-global.com/chapter/iptv-challenges-solutions-metro-networks/59677?camid=4v1a)

The Mathematics of Perception: Statistical Models and Algorithms for Image Annotation and Retrieval
[www.igi-global.com/chapter/mathematics-perception-statistical-models-algorithms/8421?camid=4v1a](www.igi-global.com/chapter/mathematics-perception-statistical-models-algorithms/8421?camid=4v1a)
Improving Auto-Detection of Phishing Websites using Fresh-Phish Framework
www.igi-global.com/article/improving-auto-detection-of-phishing-websites-using-fresh-phish-framework/196249?camid=4v1a

Hybrid Query Refinement: A Strategy for a Distance Based Index Structure to Refine Multimedia Queries
Kasturi Chatterjee and Shu-Ching Chen (2011). *International Journal of Multimedia Data Engineering and Management* (pp. 52-71).
www.igi-global.com/article/hybrid-query-refinement/58051?camid=4v1a