Introduction

The confluence of robust algorithms for structure from motion (SfM) and multi-view stereo (MVS) suggests that it will soon be feasible to accurately estimate camera pose of pictures taken in outdoor, urban environments. How can we exploit camera localization and large sets of photographs available online to better understand the contents of a particular scene? It is useful to divide scene components into static, rigid background (buildings, streets...) and dynamic objects (people, bikes, cars...).

SfM and MVS are useful techniques to build up an explicit model of the static background geometry and appearance. We want to investigate how such information can be used to improve the detection of dynamic objects like pedestrians and cars. We evaluate these ideas using a dataset of tourist photos with estimated camera pose.

Methodology

Generate 3D static scene model using internet photo collections and SfM + MVS. Use it for two approaches:

1. Scene-specific detectors utilize a stronger model of background statistics to improve accuracy
   * Unsupervised hard-negative mining via internet photo collections
   * Only need approximate camera localization at test time
2. Multi-View Background Subtraction builds a view-specific background model from nearby images
   * Estimate background mask via multi-view stereo matching
   * Suppress false-positive detections by masking out the estimated background from model

Results

Scene-specific information and geometric consistency can boost the baseline average precision up to 50% using our unsupervised negative-mining, improving Dalal-Triggs (DT) from 0.30 to 0.40 and Deformable Part Models (DPM) from 0.46 to 0.55. Using fully-supervised scene-specific negatives yielded an AP of 0.41 for DT and 0.55 for DPM, suggesting that our unsupervised negative mining is capturing most of the useful negative examples. Additionally, estimating the horizon line from SfM significantly improves the performance of related work: Putting Objects in Perspective (PoP).