Person Count Localization in Videos from Noisy Foreground and Detections

Sheng Chen[1], Alan Fern[1], Sinisa Todorovic[1]
[1]Oregon State University.

In this paper, we introduce a new problem, *person count localization* from noisy foreground and person detections. Our formulation strikes a middle-ground between person detection and frame-level counting. Given a video, our goal is to output for each frame a set of:

1. Detections optimally covering both isolated individuals and crowds of people in the video; and
2. Counts assigned to each detection indicating the number of people inside.

The problem of detecting people in videos of crowded scenes, where people frequently appear under severe occlusion by other people in the crowd is an important line of research, since detecting people in video frames has become the standard initial step of many approaches to activity recognition [1, 3, 4], and multi-object tracking by detection [6, 8, 9]. They typically use as input human appearance, pose, and orientation, and thus critically depend on robust person detections. In many domains, however, such as videos of American football or public spaces crowded with pedestrians, detecting every individual person is highly unreliable, and remains an open problem.

This motivates us to study alternative formulations that do not require perfect person localization, especially under severe clutter and occlusion, and still prove useful for higher-level video understanding. One related problem that has successfully addressed videos of crowded scenes is frame-level counting of people [2, 5, 7]. This frame-level count information, however, is a very coarse description of the video, with limited utility for high-level tasks. In particular, these problem formulations and evaluations do not address location of individuals or sub-groups.

Rather than counting people per frame, we would like to retain as much localization capability of detecting individuals as possible, but gracefully transition to counting people within areas in the frame occupied by crowds. As these crowded groups are often isolated and visually distinct from the rest of the scene, they can be viewed as "visual phrases" whose spatially tight localization and count assignment could provide useful cues for higher-level processing. For example, localized counts provide rich information about the activity unfolding during a football play by identifying many isolated and small groups of players and the primary larger player groups. Similarly, localized counts can provide space-time density statistics of crowds in an area of interest and also serve as a basis for more refined individual tracking when desired.

To solve the problem, we introduce an iterative approach called *error-driven graph revision (EGR)*, which is depicted in Fig.1. In the first iteration we extract noisy foreground objects/blobs from the video by an object detector and foreground segmentation. A corresponding initial flow graph representation G_0 that represents the temporal-spatial relationships among the foreground objects is built. An integer linear program (ILP) is then formulated based on G_0 that both selects a subset of detections and assigns counts to them, giving a solution denoted by C_0. The ILP is designed with the goal of maintaining accurate counts that also maintain temporal-spatial consistency.

At iteration i of EGR, we first look at the ILP solution C_{i-1} from the previous iteration in order to identify violations of common-sense integrity and domain constraints (for example a person cannot appear or disappear in the middle of the frame). Such violations are inevitable in our experience for any fixed way of constructing graphs from the input. Associated with each type of constraint violation are potential graph-revisions operations that may address the violation, e.g. adding edges, adding nodes, etc. A trained classifier is then used to select appropriate graph revisions to G_{i-1} that yields G_i, resulting in a new ILP and solution C_i. The iteration ends when no constraint violations are detected or a maximum number of iterations.

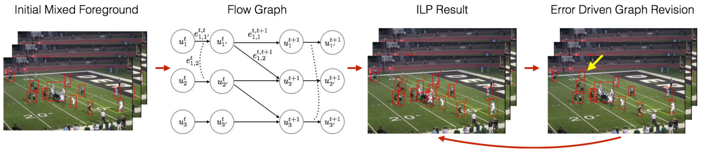

Figure 1: System Overview. First extract noisy foreground by running an object detector and foreground segmentation. A flow graph is then built and transformed into an integer program. In the following iterations, the approach detects places where integrity/domain constraints are violated by the current solution and applies one or more graph-revision operators to obtain a new graph and updated solution.

Note that, we do not assume the initial extracted foreground objects to be perfect. In fact, the iterative process aims at dealing with this noisy input. As detailed in the paper, the ILP can help address the problem of false positive foreground objects by not selecting them or assigning them counts of zero. However, the ILP does not have a natural way to deal with false negatives, which do not even appear in the corresponding flow graph. The key idea behind the EGR approach is that the ILP solutions in such cases will often violate common-sense integrity constraints that can be easily checked. Further, for a detected violation, there are natural ways to revise the graph that will potentially correct the violation, for example, by using a tracking mechanism to acquire detections in a certain space-time region of the video that were missed by the initial processing.

There are no existing metrics designed to measure the performance of count localization. Thus we propose a new metric called *count localization accuracy (CLA)* that is aimed to evaluate both count and localization accuracy. We provide experiments in two challenging domains: American football and pedestrian crowds. The results demonstrate the benefits of our approach compared to prior work and a number of baselines. Also, our evaluation suggests that EGR is a promising framework for improving other vision tasks based on fixed compilations to optimization problems.

[1] J. Aggarwal and M. Ryoo. Human activity analysis: A review. *ACM Comput. Surv*, 2011.

[2] Antoni B., Chan Zhang-Sheng John, and Liang Nuno Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *CVPR*, 2008.

[3] W. Ge, T. R. Collins, and R. B. Ruback. Vision-based analysis of small groups in pedestrian crowds. *IEEE TPAMI*, 2012.

[4] Tian Lan, Leonid Sigal, and Greg Mori. Social roles in hierarchical models for human activity recognition. In *CVPR*, 2012.

[5] V. Lempitsky and A. Zisserman. Learning to count objects in images. In *NIPS*, 2010.

[6] P. Nillius, J. Sullivan, and S. Carlsson. Multi-target tracking-linking identities using bayesian network inference. In *CVPR*, 2006.

[7] D. Ryan, S. Denman, C. Fookes, and S. Sridharan. Crowd counting using multiple local features. In *DICTA*, 2009.

[8] Guang Shu, Afshin Dehghan, Omar Oreifej, Emily Hand, and Mubarak Shah. Part-based multiple-person tracking with partial occlusion handling. In *CVPR*, 2012.

[9] Bo Yang and Ram Nevatia. Online learned discriminative part-based appearance models for multi-human tracking. In *ECCV*, 2012.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.