From Egocentric to Top-view

Shervin Ardeshir
University of Central Florida
4000 Central Florida Blvd, Orlando, FL 32816
ardeshir@cs.ucf.edu

Ali Borji
University of Central Florida
4000 Central Florida Blvd, Orlando, FL 32816
aborji@crcv.ucf.edu

Abstract

The popularity of egocentric cameras has provide us with a plethora of videos with a first person perspective. In addition, surveillance cameras and drones are rich sources of visual information, and are often captured from a top-down viewpoint. The relationship between these two very different sources of information have not been studied thoroughly and is yet to be studied. In this paper we propose to study the following problem exploring that relationship. Having a set of egocentric cameras and a top-view camera capturing the same area, we propose to identify the egocentric viewers in the top-view video. In other words, we aim to identify the people holding the egocentric cameras in the content of the top view video. For this purpose, We utilize two types of features. Unary features capturing what each individual viewer sees through time. And pairwise features encoding the relationship between the visual content of each pair of viewers. We model each view (egocentric or top) using a graph, and formulate the identification problem as an assignment problem. Evaluating our method over a dataset of 50 top-view and 188 egocentric videos taken in different scenarios demonstrates the efficiency of the proposed approach in assigning egocentric viewers to identities present in top-view camera.

1. Approach

Identifying viewers across different viewpoints could be an interesting new direction of research in computer vision. Exploring the relationships between multiple egocentric videos, or between egocentric videos and surveillance cameras could open the door to a lot of interesting research and useful applications in law enforcement and athletic events. In this effort, we attempt to address the problem of identifying egocentric viewers in a top view video. We collected a dataset containing several test cases. In each test case, multiple people were asked to move freely in a certain environment and record egocentric videos. We refer to these people as ego-centric *viewers*. At the same time, a top-view camera has recorded the entire scene including all the egocentric viewers. A more detailed version of this work has been submitted to ECCV 2016.

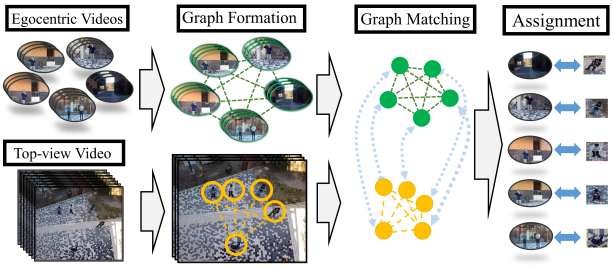

Figure 1: The input to our framework is a set of egocentric videos, and one top-view video. We aim to assign each egocentric video to one of the individuals visible in the top view video. One graph is constructed on the set of egocentric videos, where each node is an egocentric videos. Another graph is constructed on the single top-view video, where each node is an individual present in the video. We use spectral graph matching to find a soft assignment probability between the nodes of the two graphs. Using a soft-to-hard assignment, each egocentric video is matched to one of the viewers in the top-view video.

To find an assignment, each set is represented by a graph and the two graphs are compared using a spectral graph matching technique [2]. To keep track of the behavior of each individual in the top-view video, we use the multiple object tracking method proposed in [1] to compute a trajectory for each of the individuals in the top-view video. Given the fact that an egocentric video captures a person's field of view, the content of a viewer's egocentric video corresponds to the content of the individual's field of view in the top-view camera. We employ the assumption that humans mostly tend to look straight ahead, therefore having an estimate of a someone's direction of movement (which can be computed from their tracking trajectory), we can encode the changes in their field of view over time as a descriptor for each of the nodes.

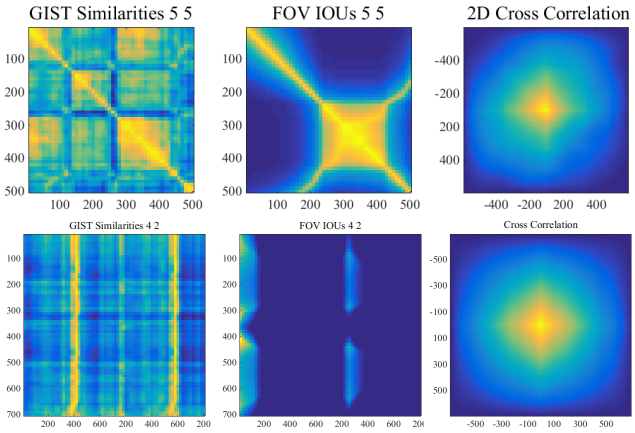

Figure 2: (a) illustrates the 2D descriptors extracted from the **nodes** of the graphs. The 2D descriptor is basically the pairwise similarities between the content of the cameras over time. Left column of (a) shows the 2D matrices extracted from the pairwise similarities of the GIST feature descriptors, middle shows the 2D descriptor capturing intersection over union of the expected FOV in the top-view camera, assuming people tend to look straight ahead. The rightmost column shows the result of the 2D cross correlation between the two, the maximum of which quantifies the similarity between the two descriptors and therefore the similarity between the two nodes. (b) shows the same concept, but between two edges. Left is the pairwise similarities between GIST descriptors of one egocentric camera to another over different time frames. Middle, is the pairwise intersection over union of the FOVs of the pair of viewers, and the rightmost column is their 2D cross correlation. The similarities between the GIST and FOV matrices in fact capture the affinity of two nodes/edges in the two graphs.

To capture a similar feature in egocentric view, We encode the changes in the global visual content (or Gist) for each of the videos, and use that as a unary feature for each node in the egocentric graph.

We also use pairwise features encoding the similarity between the content of two egocentric videos, and also the similarity between the expected content of the field of view of two viewers in the top-view camera. Examples of the descriptors can be seen in figure 2. Computing the similarity between each node/edge in the first graph to each node/edge in the second graph we can have an affinity matrix and similar to [3] compute a soft assignment from the nodes in the first graph to the nodes in the second graph. Having that matrix as the probability of each possible node to node matching, we use the well known Hungarian algorithm to come up with a hard-assignment from each egocentric video, to one of the viewers in the top view video.

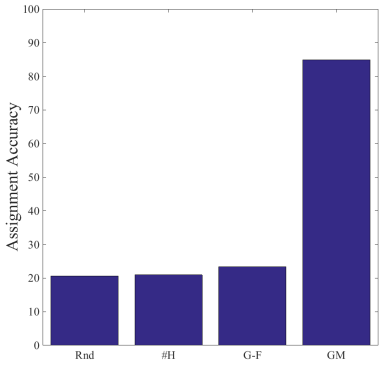

Figure 3: The assignment accuracy based on our method compared to the baselines.

2. Experimental Results

We collected a dataset containing 50 test cases of videos shot in different indoor and outdoor conditions. Each test case, contains one top-view video and several egocentric videos captured by the people visible in the top-view camera. Our dataset contains more than 225,000 frames, and the number of people visible in the top-view cameras varies from 3 to 10, while the number of egocentric cameras varies from 1 to 6. Lengths of the videos vary from 320 frames (10.6 seconds) up to 3132 frames (110 seconds).

We evaluate the accuracy of our method in terms of the percentage of the egocentric videos which were correctly matched to their corresponding viewer. The hard-assignment accuracy for our method is compared with three baselines in figure 3. Random (Rnd) in which for each egocentric video was randomly matched to one of the viewers present in the top-view video. G-F is the assignment accuracy of performing Hungarian hard assignment on the node similarities and in other words ignoring the edge similarities and the spectral graph matching step. The significant improvement of our method using both unary and pairwise features in graph matching (denoted as GM) over the baselines shows the significant contribution of pairwise features in the assignment accuracy.

References

[1] O. C. Dicle, Caglayan and M. Sznaier. The way they move: Tracking multiple targets with similar appearance. *Proceedings of the IEEE International Conference on Computer Vision*, 2013. 1

[2] Y. K. Egozi, Amir and H. Guterman. A probabilistic approach to spectral graph matching. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2013. 1

[3] R. Zass and A. Shashua. Probabilistic graph and hypergraph matching. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, June 2008. 2