Video Figure Ground Labeling

Ali Elqursh    Ahmed Elgammal

Rutgers University
{elqursh,elgammal}@cs.rutgers.edu

Abstract

Figure-ground labeling is a classical problem in computer vision in which the goal is to label different parts of the visual input as figural or background. Yet most existing approaches focus on single image figure-ground labeling with little emphasis on video. We present a method which integrates several cues to achieve figure-ground labeling on video sequences. The method is evaluated on challenging video sequences.

1 Introduction

We see the world as a 2D projection of many 3D objects on the retina. Gestalt psychologists observed that we tend to organize this clutter through a process of figure–ground segregation—i.e., by identifying those regions of the retinal images that are object-related (figures) for further processing, and relegating other regions to the background [2]. They identified many factors which play a role in identifying which regions are figure or ground. Examples of such factors include continuity, convexity, symmetry, parallelism, surroundedness, and common fate to name a few.

Although there is evidence that high-level object understanding can influence figure-ground labeling, most studies have concluded that figure-ground labeling precedes high-level processing (e.g.[8, 1]). In addition, neurological evidence suggests that motion cues play an important role in figure-ground labeling[4]. Inspired by this evidence, it seems promising to devise a method that combines low-level and mid-level cues from appearance and motion to achieve figure-ground labeling from video.

Due to the importance of figure-ground labeling in perceiving important aspects of the visual input, there exist a large literature of work on figure-ground organization. On single images, [10] proposes a shapemes descriptor to evaluate the probability of labeling a contour as belonging to figure-ground. Conditional random fields CRF are then used to infer global figure-ground assignments. [9] extends this work to integrate several low-level, mid-level, and high level cues in a single CRF formulation. However, there is no clear way to generalize such methods to video input. [13] proposes using motion cues along with appearance cues to detect occlusion boundaries. Although, figure-ground assignment of boundaries are used, such assignment do not enforce consistency or completion. In contrast, we assign labels to video segments rather than boundaries thus producing results that are always consistent. Recently, [11] proposes figure-ground labeling for egocentric videos which is a special case of our problem.

There exist a large literature on video segmentation and motion segmentation. The figure-ground labeling we address is different since video segmentation and motion segmentation attempts to segment regions that are homogeneous in color and/or motion with no notion of what is an object, figure, or ground. Similarly, the emphasis of video object segmentation is to segment out all objects without labeling them as figure or ground.

Recently there have been several attempts to extend traditional background subtraction to the moving camera settings. [12] uses orthographic motion segmentation over a sliding window to segment a set of trajectories. This is followed by sparse per frame appearance modeling to densely segment images. In [7], an iterative method is proposed that maintains block based appearance models in a Bayesian filtering framework. Since such methods typically use dominant motion or occlusion cues to determine what is foreground and background, they do not always match our expectations.
about what is figure or ground.

In this paper we propose a method that achieves figure-ground labeling from video sequences. We combine saliency, motion, and color features into a single energy function which we optimally optimize. By using saliency, our labelings more closely match what we typically consider figure and at the same time combines many low-level cues in the process. On the other hand, motion and color similarity help disambiguate figure and ground in regions where the saliency cue is least confident.

2 Approach

Our approach consists of several stages; see Figure 2. In the first stage, we preprocess the video by applying video segmentation and dense trajectory tracking (Subsection 2.1). Next we compute a saliency, motion, and color features for each video segment (Subsection 2.2). Finally, we formulate an energy function that integrates saliency, motion and color features and find the optimal assignment of figure-ground labels by minimizing it (Subsection 2.3).

2.1 Preprocessing

In order to obtain robust features, we aggregate information over a large number of voxels in the video. This is achieved by applying the hierarchical video segmentation approach of [3] to efficiently obtain a video segmentation. The result of this step is a set of segments $V = \{V_1, \ldots , V_N\}$ with each segment $V_i$ represented with a set of video voxels $V_i = \{(x, y, t)\}$.

Next, in preparation to computing motion features, we extract dense trajectories using large displacement optical flow (LDOF) [14]. Denoting trajectory $j$ by $T_j$, we assign each trajectory to a segment $\nu(T_j)$ by first finding the segment to which each point along the trajectory belongs to and then selecting the mode of this list.

2.2 Feature Extraction

Rather than computing many local feature and determine what features are discriminative for figure-ground labeling, we use state of the art saliency method of [5] to integrate many low and mid-level image features and produce a saliency map. Such maps are computed for each individual frame and then used to compute a saliency feature for each video segment. Let $S$ represent the saliency volume. We define the saliency of a segment $s_i$ as the mean saliency of all the voxels belonging to the segment

$$s_i = \frac{\sum_{(x, y, t) \in V_i} S(x, y, t)}{|V_i|}.$$

Saliency maps are relatively noisy and without considering the relations between video segments, figure-ground labeling is inaccurate at regions where saliency is least confident. Therefore, we compute color features and motion features for each video segment and use them to measure the similarity between any pair of video segments.

Since video segments represent regions with homogeneous color, we use the mean color $c_i$ of a segment as a color feature. To compute it we first convert the colors to Lab color-space and then compute the mean over all voxels belonging to each segment. The Lab color-space is robust to changes in lightness and is able to capture chromatic similarity under a wide range of lighting conditions.

$$c_i = \frac{\sum_{(x, y, t) \in V_i} c(x, y, t)}{|V_i|}.$$

To compute the motion features $m_i$ we first transform each trajectory into a vector of relative motions between frames. This effectively removes the dependency on the starting location. Next for each segment we compute the mean relative motion vector $m_i$. Since for some segments we may have no trajectories occupying a pair of frames we define $o_i$ as an indicator vector such that element $o_{ij} = 1$ if and only if there exist trajectories in segment $i$ overlapping frames $j$ and $j + 1$.

2.3 Figure-Ground Labeling

To incorporate the pairwise relations between the segments, we first define a graph structure over the segments. With each node representing a segment $V_i$ we define the set of neighbors $\mathcal{N}(V_i)$ as the set of segments which shares any boundary with $V_i$. However, it is not infrequent that a ground region is surrounded by figure or vice versa. To solve this we augment the set of neighbors with other segments which are similar in color. Formally, we cluster the color features from all the segments into $k$ clusters using k-means. For each segment in a cluster we augment its neighbors with other segments in the same cluster.

To achieve figure-ground labeling we look for a labeling $L = [l_1 \ldots l_N]$ that satisfies the following criteria. First, figure is salient while ground is non-salient. Second, the appearance and motion of neighboring segments with the same label vary smoothly. We formulate an energy function which encapsulates this criteria as

$$E(L) = \sum_{V_i \in \mathcal{V}} |\mathcal{N}(V_i)| f(l_i) + \sum_{V_i} \sum_{V_j \in \mathcal{N}(V_i)} g(l_i, l_j),$$

where $f$ is a unary term which measures how well the label satisfies the saliency feature of the segment, and $g$ is a smoothness term that measures compatibility of
segments with the same label. By weighting with the number of neighbors we are able to adaptively weight the unary term in such a way that it avoids the bias due to a large number of neighbors. The unary potential $f$ is defined as

$$f(l_i) = (-l_is_i + (1 - l_i)(1 - s_i)).$$

Here we penalize low saliency for figure segments ($l_i = 1$) and high saliency for ground ($l_i = 0$). The smoothness term $g$ is a combination of two terms $g_c$, and $g_m$ which measure color and motion compatibility respectively

$$g(l_i, l_j) = \lambda g_c(l_i, l_j) + (1 - \lambda)g_m(l_i, l_j).$$

$\lambda$ is a parameter that measures the relative importance of each term. The terms $g_c$, and $g_m$ are defined as

$$g_c(l_i, l_j) = \begin{cases} 0 & l_i \neq l_j \\ \exp(-\frac{1}{2}(c_i - c_j)^T\Sigma_c^{-1}(c_i - c_j)) & l_i = l_j \end{cases},$$

where $\Sigma_c$ is a diagonal covariance matrix.

$$g_m(l_i, l_j) = \begin{cases} 0 & l_i \neq l_j \\ \exp(-\frac{\|m_i - m_j\|_2}{2(\sigma_i^2 \sigma_j^2)}) & l_i = l_j \end{cases},$$

where $\sigma_m$ is a parameter that controls how strongly dissimilarity is penalized. Both $g_c$ and $g_m$ penalizes large color or motion differences between neighboring segments with the same label.

The global minimum of the energy function (1) can be found using graph cuts [6]. The obtained labels induces a final labeling $F = \cup_{l_i=1}^{V_i}$.

3 Evaluation

We evaluate the approach on videos from the occlusion boundary detection dataset from [13]. It consists of 30 short video sequences (approximately 10-20 frames each). Each exhibits very brief camera motion, instantaneous motion of objects in the scene, or a combination of the two. However, since the dataset was created with the emphasis of boundary detection, the ground truth provided is for individual object segmentations and not for figure-ground segmentation. We resort to qualitative evaluation of the method on these sequences. To evaluate the approach on rapid motion we also use the *yunakin* sequence from the youtube dataset [3]. For all sequences we choose the following values for our parameter settings $\Sigma_c = diag(100, 15, 15)$, $\lambda = \frac{1}{2}$, $\sigma_m = 2$, $k = 5$.

To our knowledge there exist no prior work on figure-ground labeling for videos in the general setting. To demonstrate the efficacy of our approach, we compare our approach with a baseline which does not use motion or color features. Instead, it uses a smoothness prior that encourages spatial smoothness. Figure 3 shows the results on 3 sequences (*rocking horse, walking legs, and post*) from [13] and the *yunakin* sequence from youtube dataset [3]. For each video we show one frame from the sequence, the video segmentation result, color features, saliency features, result of the baseline, and the result of the approach.

Comparing the results of the baseline and the approach shows that the baseline suffers from missing parts (*rocking horse*), no figure (*walking legs, post*), or no ground (*yunakin*). It indicates that saliency information is not enough to determine figure-ground assignment of video segments. This can be explained in part by the fact that saliency gives a rough estimate of where attention is focused in the image and does not define a grouping of regions into figure. In addition, regions close to object boundaries are usually affected by the existence of a salient region besides it. Our approach is able to successfully obtain the correct labeling in all of the sequences. Errors in the results can be attributed to inaccurate video segmentation along the boundary (*walking legs, post*), or small segments close to the object that has a high saliency feature (*walking legs*).
4 Conclusion

We presented an approach that accurately computes figure-ground labelings from video sequences. By leveraging saliency information together with color and motion cues, it produces labelings that encapsulate salient regions while respecting the natural grouping of objects together. We demonstrated its efficacy by evaluating it on challenging sequences that contain both non-rigid and fast motion. In the future we would like to create a benchmark video dataset for figure-ground labeling of video sequences based on user input.

References