Automatic discovery of foreground objects in video sequences is important in computer vision, with applications to object tracking, video segmentation and weakly supervised learning. This task is related to cosegmentation [4, 5] and weakly supervised localization [2, 6]. We propose an efficient method for the simultaneous discovery of foreground objects in video and their segmentation masks across multiple frames. We offer a graph matching formulation for bounding box selection and refinement using second and higher order terms. It is based on an Integer Quadratic Programming formulation and related to graph matching and MAP inference [3]. We take into consideration local frame-based information as well as spatiotemporal and appearance consistency over multiple frames. Our approach consists of three stages. First, we find an initial pool of candidate boxes using a novel and fast foreground estimation method in video (VideoPCA) based on Principal Component Analysis of the video content. The output of VideoPCA combined with Edge Boxes [8] is then used to produce high quality bounding box proposals. Second, we efficiently match bounding boxes across multiple frames, using the IPFP algorithm [3] with pairwise geometric and appearance terms. Third, we optimize the higher order terms using the Mean-Shift algorithm [1] to refine the box locations and establish appearance regularity over multiple frames. We make the following contributions:

1. A novel formulation with efficient discrete and continuous optimization for joint selection and refinement of object bounding boxes in video. Our approach encourages appearance, geometric and spatiotemporal consistency over multiple frames with a formulation that considers relations between neighboring as well as farther away frames (Figure 1).

2. A fast method for estimating foreground regions based on Principal Component Analysis of the video content that is able to handle cases of moving or changing backgrounds.

Given a video \( V \) as a sequence of temporally ordered frames \( V = \{I_1, I_2, \ldots, I_t\} \), the goal is to find the main foreground object present in the sequence. For each frame \( I_i \) we have a pool of \( n_i \) candidate bounding boxes \( B_{ia} \)'s, obtained automatically. For each box \( a \) we store its \( xy \) location in image \( i \) in \( \theta_{ia} \). Let \( x_{ia} \) be an indicator variable corresponding to the bounding box \( B_{ia} \) (the a-th bounding box of frame i), such that \( x_{ia} \) is 1 if the bounding box \( B_{ia} \) is selected, and 0 otherwise. The indicator variables are arranged in vector form \( x \) such that its \( ia \)-th element corresponds to \( x_{ia} \). We impose the constraint that a single box can be selected per frame: \( \sum_a x_{ia} = 1 \). Thus, vector \( x \) represents a discrete solution that indicates which box is selected. Similarly, we keep the continuous location parameters in a global vector \( \theta \), with \( \theta_{ia} \) being the parameters of box \( B_{ia} \). The problem becomes one of joint box matching and location refinement over multiple frames, in which we optimize over both \( x \) and \( \theta \):

\[
(x^*, \theta^*) = \arg\max_{x, \theta} (x^T M x + \sum_{i=1}^n H_i(x, \theta))
\]

\[
s.t. \sum_i x_{ia} = 1 \ \forall i, \ x \in \{0,1\}^n.
\]

We adopt a two stage approach. First stage performs discrete optimization in which the quadratic function is optimized by finding the correct frame-to-box matches \( x \), given a fixed \( \theta \). The second stage, when \( x \) is fixed, the location \( \theta \) is refined by the non-parametric Mean-Shift to locally optimize the foreground pixels density.

The pairwise terms have the following form:

\[
M_{ia,jb} = \exp(w^T g_{ia,jb}),
\]

such that: 1) \( f_{ia} \) and \( f_{jb} \) respectively measure the absolute difference between the average foreground soft-segmentation values in box \( a \) (and \( b \) respectively) vs. average foreground values in the surrounding background from its frame \( i \) (and \( j \) respectively). 2) Similarly, \( v_{ia} \) and \( v_{jb} \) measure the absolute difference in the relative mean speed between the box and the surrounding background, computed using the DeepFlow method [7]. 3) \( c_{ia} \) and \( c_{jb} \) measure the distance between the box center and the image center. 4) \( m_{ia,jb} \) reflects the quality of the match between the standard HOG descriptor of box \( a \) and that of box \( b \). 5) \( r_{ia,jb} \) measures the overlap-over-union between the boxes. 

For the higher order terms we estimate foreground and background color probability distributions from the \( 2k+1 \) bounding boxes and their frames connected to the current one (including itself) and estimate the soft foreground segmentation by classifying individual pixels based on their color. The higher order term \( H_i(x, \theta) = A_{cl}(i) \) measures the difference between average foreground segmentation values inside the box and outside of it.

We test our method on the large scale YouTube-Objects dataset and obtain state-of-the-art results on several object classes, at a low computational cost. We conclude that our approach, by proposing efficient bounding box generation with VideoPCA, fast matching over multiple frames, and high quality soft-segmentation estimation, covers and extends current approaches in object discovery in video.

\[\text{Figure 1: The structure of our box-matching formulation allows links between neighboring frames as well as farther away ones.}\]