Player segmentation in team sports videos is crucial to video semantic analysis, such as player interaction identification and tactic analysis. We leverage the appearance similarity among players of the same team, and cast this task as a co-segmentation problem. In this way, the extra knowledge shared across players significantly reduces unfavorable uncertainty in segmenting individual players. We are also aware that the performance of co-segmentation highly depends on the used features, and further propose a contrast-based approach to estimate the discriminant power of each feature in an unsupervised manner. It turns out that our approach can properly fuse features by assigning higher weights to discriminant ones, and result in remarkable performance gains. The promising results on segmenting basketball players manifest the effectiveness of our approach.

Index Terms— Player segmentation, sports video understanding, co-segmentation, contrast-aware feature selection

2. RELATED WORKS

Sports video analysis has attracted interest and been explored for a long time. It spreads a wide spectrum of issues. Liu et al. [5] inferred shot and scene segmentation by using motion information. Han et al. [6] analyzed camera movement by referring to the motion vector for game state estimation. Peršć et al. [7] categorized team activities by detecting the positions and trajectories of players. Chen et al. [8]
instead detected ball trajectories and shooting positions. Bailley and Jose [9] investigated the audio part of sports videos to identify key events. In the works by Liu et al. [10] and Zhang et al. [11], visual, audio, and motion cues were jointly considered to bridge the gap between broadcast videos and play-by-play texts. Lu et al. [1] combined three visual information sources for player identification, including raw image, MSER [12] visual words, and SIFT [13] visual words. In the aforementioned works, we are aware of a research trend where intra- and inter-player analysis are emphasized. Hence, accurate and efficient player segmentation gradually becomes essential to nowadays sports video analysis.

Co-segmentation, firstly introduced by Rother et al. [4], aims to simultaneously segment the common foregrounds of multiple images. A vast amount of recent research efforts has made significant progress of co-segmentation. One branch of approaches to image co-segmentation is based on Markov random field (MRF). Rother et al. [4] employed an MRF model over images, and enforced a global consistency term among foreground histograms. Yu et al. [14] and Chang et al. [15] incorporated the co-saliency prior into co-segmentation for foreground identification. Hochbaum and Singh [2] used rewarded affinities instead of penalty terms to better solve MRF optimization. Another line of co-segmentation methods is based on graph-partitioning. Joulin et al. [16] merged bottom-up image segmentation and top-down class separation into a unified discriminative graph matrix, and derived the figure-ground labels by graph partitioning. Joulin et al. [17] further generalized their work to multi-class co-segmentation. Kim et al. [18] applied hierarchical clustering to image grouping, compiled multiple levels of segmentation, and used intra and inter-image connections to carry out co-segmentation. In this work, we cast the task of player segmentation as a co-segmentation problem upon Joulin et al.’s model [16], and further improve its performance via adaptive feature selection.

Information fusion is referred to as the integration of multiple media, features, or intermediate decisions. It serves as a feasible way for improving performance. Atrey et al. [19] summarized many existing approaches to information fusion, and categorized them into three groups according to the levels of fusion, i.e., feature level, decision level, and hybrid. Our method belongs to feature-level fusion. Since image co-segmentation is an unsupervised task, most supervised feature selection and fusion algorithms are not applicable. We instead assess the discriminative power of each feature according to the divergence of that feature’s responses inside and outside the bounding boxes of players. Then, we dynamically generate proper weights of all the features for their combination.

3. OUR PROPOSED APPROACH

In this section, we firstly describe how to formulate player segmentation as a co-segmentation problem. Then we introduce the co-segmentation algorithm by Joulin et al. [16] upon which our approach is conducted, and show how to improve its performance by adaptive feature fusion.

3.1. Problem statement

Considering a frame of a sports video where \( m \) players of the same team present, our goal is to segment these players as precisely as possible. Assume the bounding boxes \( \{ B_i \}_{i=1}^m \) of the \( m \) players are given in advance, say by using an off-the-shelf detector. The implicit segments \( \{ C_i \}_{i=1}^m \) of the \( m \) players can then be estimated by using any segmentation algorithm. However, most segmentation algorithms suffer from various difficulties in this application, such as cluttered backgrounds on the court, and moving and non-rigid players. We observe in Fig. 1 that the foreground areas within the bounding boxes are highly consistent owing to the common uniforms and similar player skins, while the background areas instead exhibit diversity, such as audiences and floor boards. This observation allows us to formulate the task of seeking \( \{ C_i \}_{i=1}^m \) as an image co-segmentation problem by taking \( \{ B_i \}_{i=1}^m \) as input. We call this new task as single-frame co-segmentation, since all regions to be segmented come from a single frame. Compared with conventional co-segmentation, single-frame co-segmentation gives extra information. The region outside all the bounding boxes provides the prior knowledge about the backgrounds within the bounding boxes. We utilize this property to identify good features for co-segmentation.

3.2. Co-segmentation algorithm by Joulin et al. [16]

The literature on image co-segmentation is quite extensive. Our approach is established upon the discriminative clustering algorithm by Joulin et al. [16], because it considers both inter-image similarity and intra-image spatial consistency, and achieves the state-of-art performance. With input bounding boxes \( \{ B_i \}_{i=1}^m \), Joulin et al.’s algorithm partitions pixels in \( \{ B_i \}_{i=1}^m \) into foregrounds and backgrounds, and represents the results by \( y = [y_1^T, y_2^T, \cdots, y_m^T]^T \in \{-1, 1\}^n \), where \( y_i \in \{-1, 1\}^n \) is the figure-ground separation of \( B_i \), \( n_i \) is the number of pixels in \( B_i \), and \( n = \sum_{i=1}^m n_i \). The co-segmentation model [16] employs discriminative matrix \( A \in \mathbb{R}^{n \times n} \) and spatial consistency matrix \( L \in \mathbb{R}^{n \times n} \), and infers co-segmentation results \( y \) by solving the following constrained optimization problem:

\[
\begin{align*}
\min_y & \quad y^T \left( A + \frac{\mu}{n} L \right) y \\
\text{s.t.} & \quad \forall B_i, \quad \lambda_0 n_i \delta_i \leq \frac{1}{2} (yy^T + 1_n 1_n^T) \delta_i \leq \lambda_1 n_i \delta_i,
\end{align*}
\]

where \( \delta_i \in \{0, 1\}^n \) is the indicator vector of \( B_i \) with \( (\delta_i)_j = 1 \) if the \( j \)th pixel belongs to \( B_i \) and 0 otherwise. \( \lambda_0 \) and \( \lambda_1 \) represent the lower bound and the upper bound of the cluster size, respectively. In our case, the foreground, i.e., player, in a bounding box is neither too large nor too small, so we set
\(\lambda_0 = 0.2\) and \(\lambda_1 = 0.8\). Parameter \(\mu\) controls the tradeoff between bottom-up segmentation and discriminative clustering. We empirically set \(\mu\) as 0.1.

With discriminative matrix \(A\), the first term, \(y^T Ay\), of the objective function in (1) represents the separability of the predicted foreground and background. This term is derived based on the loss function parameterized by a kernel matrix, which takes all pixels across different bounding boxes into account. Minimizing the loss function improves the separability of the foreground and background across boxes. The second term, \(y^T Ly\), encodes both the visual (color) and spatial similarity between pixels residing in the same bounding box, and enforces intra-box consistency in co-segmentation. Specifically, \(L\) is the graph Laplacian of affinity matrix \(W \in \mathbb{R}^{n \times n}\). \(W\) is a block-diagonal matrix by assembling separate similarity \((W_i \in \mathbb{R}^{n_i \times n_i}\), \(i=1\) on the diagonal. Each \(W_i = [W_{uv}]\) can be further separated into color similarity \(W_{uv,c} = [W_{uv,c}]\) and location similarity \(W_{uv,p} = [W_{uv,p}]\) for a pair of pixels \(u\) and \(v\) in box \(i\). Their definitions are given by

\[
W_{uv} = W_{uv,c} \times W_{uv,p}
\]

\[
= \begin{cases} 
\exp \left( -\|e_u - e_v\|^2 - \lambda \|p_u - p_v\|^2 \right), & \|u - v\| \leq 2, \\
0, & \text{otherwise},
\end{cases}
\]

where \(p_u = [p_{u, r}, p_{u, g}, p_{u, b}]^T \in \mathbb{R}^2\) and \(e_u = [e_{u, r}, e_{u, g}, e_{u, b}]^T \in \mathbb{R}^3\) are the 2D coordinate and the RGB color of pixel \(u\) respectively, and \(\lambda\) is a positive constant controlling the tradeoff between the color and spatial evidences.

3.3. Adaptive feature weighting and fusion

The performance of co-segmentation highly relies on the selected features. The color evidences in this application are quite important. However, the relative importance among the \(R\), \(B\), and \(G\) channels typically varies from video to video, even from frame to frame. It depends on the colors of uniform, the court, and so on. In viewing of this property, we change the color-based affinity matrix \(W_c\) from (2) to

\[
W_{uv,c} = \begin{cases} 
\exp \left( - \sum_{f \in \{r, g, b\}} w_f \|c_{uf} - c_{vf}\|^2 \right), & \|u - v\| \leq 2, \\
0, & \text{otherwise},
\end{cases}
\]

where \(f\) is the index of color channels, each of which is associated with weight \(w_f\). Color channels with higher weights have larger impact on the results of co-segmentation. Channels that are discriminative between foreground and background are considered more important, and should be associated with higher weights. It leads to the cause-and-effect dilemma for jointly solving co-segmentation and seeking feature weights, since we know neither foreground-background separation nor the optimal features in advance in the unsupervised co-segmentation task.

Since the true contour of the player within a bounding box is unknown, we can’t compute the contrast of a feature between the true foreground and background. Thus, we consider an alternative contrast. For each bounding box \(B_i\), we enlarge it by a certain margin so that the enlarged one is twice as large as the original one. As shown in Fig. 2, bounding box \(B_i\) contains both the foreground (player) and the background (court), while the region between the two bounding boxes, denoted by \(B_i'\), covers only the background (Of course, we exclude the region that overlaps another player). It follows that we can estimate the effectiveness of a color feature according to the diversity between its responses in \(B_i\) and \(B_i'\).

Specifically, for each color channel \(f\), we quantize the response range of this channel into \(D\) bins, and compile two histograms, one for \(B_i\) and one for \(B_i'\), based on the quantized pixel responses. Denote the two histograms as \(x = [x_1, x_2, \ldots, x_D]\) and \(x' = [x_1', x_2', \ldots, x_D']\), respectively. The contrast of color channel \(f\) is measured by using \(\chi^2\) distance, i.e.,

\[
\chi^2_f = \frac{1}{2D} \sum_{d=1}^{D} \frac{(x_d - x'_d)^2}{(x_d + x'_d)}. \tag{4}
\]

The larger the \(\chi^2_f\) is, the more discriminative the color \(f\) is. Thus, we define the channel weight \(w_f\) in (3) as

\[
w_f = \chi^2_f / \tau, \tag{5}
\]

where \(\tau\) is a positive constant, and is empirically set as 9 in all the experiments. By adaptively computing the contrast of the \(R\), \(G\), and \(B\) color channels, we put higher weights on the channels that lead to better separation between the foreground and background. As shown in the experiments, the yielded co-segmentation results are remarkably improved.

4. EXPERIMENTAL RESULTS

Our approach is evaluated on basketball games. We assume that the bounding boxes of the players are available by either manual labeling or using an existing detector. In the experiments, we choose to manually and precisely label them for the sake of evaluation so that the induced errors are then totally caused by the segmentation algorithms. Specifically, we select eight frames from NBA 2015 playoff first round, including four from home teams and four from guest teams. Each frame contains the five players of a team and has no significant mutual occlusion among the players.
We compare our method to traditional spectral clustering by using the implementation in [20], which applies normalized cut to each bounding box individually. The other method for comparison is the discriminative clustering algorithm by Joulin et al. [16], which like our method, takes the five bounding boxes of the players in a frame into account jointly. We set the target number of segments as two, i.e., foreground and background, for our method and the two compared methods.

In order to compare the three methods quantitatively, we adopt segmentation precision, i.e., intersect over union (IoU), as the evaluation metric:

$$\text{Precision} = \frac{\text{GT} \cap P}{\text{GT} \cup P},$$

where $\text{GT}$ stands for the ground truth of a player, while $P$ is the predicted segment by a segmentation algorithm. Note that each of our method and the two compared methods partitions a bounding box into two segments. In the unsupervised setting, we pick the one with the higher precision in (6) as the foreground, and report the performance.

By averaging over all the bounding boxes in the eight frames, the performance, in precision, of our approach and the two compared ones is reported in Table 1. It can be observed that co-segmentation gives much higher performance than individual segmentation. It confirms that the consistence between foregrounds (players) of a team is an important clue to alleviate the difficulties in the challenging segmentation tasks. Our approach further introduces adaptive feature selection into co-segmentation, and achieves superior results.

Table 1. Precision of segmentation methods in $[\text{mean} \pm \text{std}]$.

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spectral Clustering [20]</td>
<td>0.41 ± 0.18</td>
</tr>
<tr>
<td>Co-segmentation [16]</td>
<td>0.47 ± 0.18</td>
</tr>
<tr>
<td>Ours</td>
<td>0.51 ± 0.19</td>
</tr>
</tbody>
</table>

The co-segmentation framework enforces the consistence of the common foregrounds, and hence can better separate the players from the basketball courts. Our method measures the discriminative powers of the R, G, and B channels, and further improves the co-segmentation results by putting emphasis on more discriminative color channels.

5. CONCLUSIONS

We have addressed the challenging task of player segmentation in team sports videos. Unlike conventional approaches that conduct individual player segmentation, we reformulate it as a single-frame co-segmentation task, and illustrate it upon the state-of-art co-segmentation framework by leveraging the properties of team sports. Motivated by the lack of a systematic way for feature selection in conventional co-segmentation methods, an algorithm is presented to estimate the discriminative power of each feature, and adaptively associate these features with proper weights for their fusion. We have shown that our proposed approach can enhance segmentation performance on challenging team sports videos in the experiments. It is worth mentioning that our approach doesn’t make use of sport-specific properties. For future work, we will put emphasis on applying the approach to various team sports videos, such as tennis, volleyball, and soccer.

Acknowledgement. This work was supported by Ministry of Science and Technology (MOST) under grants 103-2221-E-001-026-MY2, 104-2628-E-001-001-MY2, and 103-2221-E-001-009-MY3.
6. REFERENCES


