PRECISE PLAYER SEGMENTATION IN TEAM SPORTS VIDEOS USING CONTRAST-AWARE CO-SEGMENTATION

Tsung-Yu Tsai[⋆†] *Yen-Yu Lin*[⋆] *Hong-Yuan Mark Liao*[⋆] *Shyh-Kang Jeng*[†]

[⋆]Academia Sinica [†]National Taiwan University

ABSTRACT

Player segmentation in team sports videos is challenging but crucial to video semantic understanding, 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

1. INTRODUCTION

Player segmentation in team sports videos is crucial to video semantic analysis, such as player pose estimation, interaction recognition, and tactic analysis, since rich visual evidences inferred from play contours facilitate these tasks. This task is typically cast as an object segmentation problem, which is widely studied, but still remains challenging in general. The situation becomes even more difficult for segmenting players in a team sports video. Take the half-court view of a basketball game in Fig. 1 as an example. The cluttered background on the court, moving and non-rigid players, mutual occlusions often lead to unsatisfactory segmentation results.

Player segmentation is difficult. Nevertheless, player detection is relatively easy, and has been well solved by powerful detectors, such as [1]. It can be observed in Fig. 1 that the bounding boxes of the detected players of a team share highly similar appearances in the foreground areas, i.e., players, while have diverse backgrounds. With a player detector and this prior observation, we propose to formulate team sports player segmentation as a *co-segmentation problem* [2, 3, 4]. In this way, the extra knowledge transferred from other players can be utilized to reduce the complexity of player segmentation. Unlike most co-segmentation problems, the targets for co-segmentation in our cases are the detected bounding boxes in a single image. Our approach to player segmentation is illustrated for basketball games in this paper, but we consider it general enough to be applied to many other team sports, such as hockey, football, and volleyball.

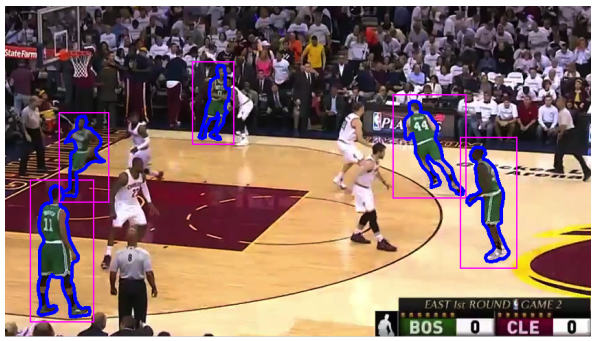

Fig. 1. The half-court view of a basketball game. The players of a team are detected in magenta bounding boxes. Our goal is to find the blue contours of these players.

The performance of co-segmentation extremely depends on the adopted features, but its unsupervised nature makes feature selection almost infeasible. Nevertheless, the bounding boxes available in our cases reveal the clues for estimating the goodness of a feature. We consider that a feature is effective for segmentation if it is *discriminative* enough to distinguish foregrounds from backgrounds. Specifically, for each feature, we evaluate its *contrast* between the regions outside and inside the bounding boxes, and then develop an algorithm for adaptive feature fuse. It turns out that more discriminative features are selected to provide better figure-ground separation, and result in remarkable performance boost.

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. Perše 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. Baillie 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 $y = [y_1^\top \; y_2^\top \; \cdots \; y_m^\top]^\top \in \{-1, 1\}^n$, where $y_i \in \{-1, 1\}^{n_i}$ 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:

$$\begin{aligned} &\min y^\top \left(A + \tfrac{\mu}{n} L\right) y \\ &\text{s.t. } \forall B_i,\ \lambda_0 n_i \delta_i \le \tfrac{1}{2} \left(y y^\top + 1_n 1_n^\top\right) \delta_i \le \lambda_1 n_i \delta_i, \end{aligned} \tag{1}$$

where $\delta_i \in \{0, 1\}^n$ is the indicator vector of B_i with $(\delta_i)_j = 1$ if the jth pixel belongs to B_i and 0 otherwise. λ_0 and λ_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 μ controls the tradeoff between bottom-up segmentation and discriminative clustering. We empirically set μ as 0.1.

With discriminative matrix A, the first term, $y^\top Ay$, 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, $y^\top Ly$, 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 $\{W^i \in \mathbb{R}^{n_i\times n_i}\}_{i=1}^m$ on the diagonal. Each $W^i = [W^i_{uv}]$ can be further separated into color similarity $W^i_c = [W^i_{uv,c}]$ and location similarity $W^i_p = [W^i_{uv,p}]$ for a pair of pixels u and v in box i. Their definitions are given by

$$W^i_{uv} = W^i_{uv,c} \times W^i_{uv,p} = \begin{cases} \exp\left(-\|\mathbf{c}^u - \mathbf{c}^v\|^2 - \lambda\|\mathbf{p}^u - \mathbf{p}^v\|^2\right), & \|u-v\| \le 2, \\ 0, & \text{otherwise}, \end{cases} \quad (2)$$

where $\mathbf{p}^u = [p^u_x,\ p^u_y]^\top \in \mathbb{R}^2$ and $\mathbf{c}^u = [c^u_r,\ c^u_g,\ c^u_b]^\top \in \mathbb{R}^3$ are the 2D coordinate and the RGB color of pixel u respectively, and λ 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 adopted 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 uniforms, the court, and so on. In viewing of this property, we change the color-based affinity matrix W^i_c from (2) to

$$W^i_{uv,c} = \begin{cases} \exp\left(-\sum_{f\in\{r,g,b\}} w_f\|c^u_f - c^v_f\|^2\right), & \|u-v\| \le 2, \\ 0, & \text{otherwise}, \end{cases} \quad (3)$$

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 $B_i^{'}$.

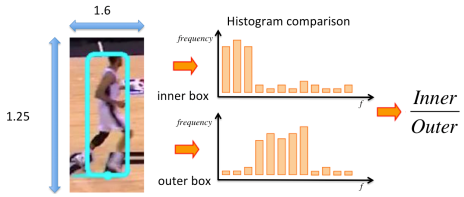

Fig. 2. Original bounding box (cyan) and the enlarged one. The effectiveness of a color feature can be estimated by the *contrast* between the feature responses of the two boxes.

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 according quantized pixel responses. Denote the two histograms as $\mathbf{x} = [x_1,\ x_2,\ \ldots,\ x_D]$ and $\mathbf{x}^{'} = [x_1^{'},\ x_2^{'},\ \ldots,\ x_D^{'}]$, respectively. The contrast of color channel f is measured by using χ^2 distance, i.e.,

$$\chi^2_f = \frac{1}{2D}\sum_{d=1}^{D}\frac{(x_d - x_d^{'})^2}{(x_d + x_d^{'})}. \quad (4)$$

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

$$w_f = \chi^2_f/\tau, \quad (5)$$

where τ 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 are 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.

Fig. 3. (A) and (B) Two examples of the segmentation results. (a) The five bounding boxes of the players in a team. Segmentation results by (b) spectral clustering [20], (c) co-segmentation algorithm [16], and (d) our approach.

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:

$$\text{Precision} = \frac{GT \cap P}{GT \cup P}, \tag{6}$$

where 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.

To gain insight into the quantitative results, we show two examples in Fig. 3. As we can see, because players' uniforms may be similar to the basketball court in color, the results by spectral clustering for individual player segmentation [20] are not satisfactory. 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.

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

Method	Precision rate
Spectral Clustering [20]	0.41 ± 0.18
Co-segmentation [16]	0.47 ± 0.18
Ours	0.51 ± 0.19

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

[1] W.L. Lu, J.A. Ting, J.J. Little, and K.P. Murphy, "Learning to Track and Identify Players from Broadcast Sports Videos," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2013.

[2] D.S. Hochbaum and V. Singh, "An efficient algorithm for Co-segmentation," in *Proc. Int'l Conf. Computer Vision*, 2009.

[3] Y. Mu and B. Zhou, "Co-segmentation of image pairs with quadratic global constraint in mrfs," in *Proc. Asian Conf. on Computer Vision*, 2007.

[4] C. Rother, T. Minka, A. Blake, and V. Kolmogorov, "Cosegmentation of image pairs by histogram matching-incorporating a global constraint into MRFs," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2006.

[5] S. Liu, H. Yi, L.T. Chia, D. Rajan, and S. Chan, "Semantic analysis of basketball video using motion information," in *Proc. Pacific Rim Conf. on Multimedia*, 2004.

[6] X. Han, L. Wu, X. Liu, Z. Cheng, and Y. Gong, "Offense-defense semantic analysis of basketball game based on motion vector," in *Proc. Int'l Conf. Image Analysis and Signal Processing*, 2009.

[7] M. Perše, M. Kristan, S. Kovačič, G. Vučkovič, and J. Perš, "A trajectory-based analysis of coordinated team activity in a basketball game," *Computer Vision and Image Understanding*, 2009.

[8] H.-T. Chen, M.-C. Tien, Y.-W. Chen, W.-J. Tsai, and S.-Y. Lee, "Physics-based ball tracking and 3d trajectory reconstruction with applications to shooting location estimation in basketball video," *J. Visual Communication and Image Representation*, 2009.

[9] M. Baillie and J.M. Jose, "An audio-based sports video segmentation and event detection algorithm," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2004.

[10] S. Liu, M. Xu, H. Yi, L.T. Chia, and D. Rajan, "Multimodal semantic analysis and annotation for basketball video," *EURASIP J. on Applied Signal Processing*, 2006.

[11] Y. Zhang, C.S. Xu, Y. Rui, J. Wang, and H. Lu, "Semantic event extraction from basketball games using multi-modal analysis," in *Proc. Int'l Conf. Multimedia and Expo*, 2007.

[12] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *J. Image and Vision Computing*, 2004.

[13] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int'l J. Computer Vision*, 2004.

[14] H. Yu, M. Xian, and X. Qi, "Unsupervised co-segmentation based on a new global gmm constraint in mrf," *Proc. Int'l Conf. Image Processing*, 2014.

[15] K.-Y. Chang, T.-L. Liu, and S.-H. Lai, "From co-saliency to co-segmentation: An efficient and fully unsupervised energy minimization model," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2011.

[16] A. Joulin, F. Bach, and J. Ponce, "Discriminative clustering for image co-segmentation," *Proc. Conf. Computer Vision and Pattern Recognition*, 2010.

[17] A. Joulin, F. Bach, and J. Ponce, "Multi-class cosegmentation," *Proc. Conf. Computer Vision and Pattern Recognition*, 2012.

[18] E. Kim, H. Li, and X. Huang, "A hierarchical image clustering cosegmentation framework," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2012.

[19] P.K. Atrey, M.A. Hossain, A. El Saddik, and M.S. Kankanhalli, "Multimodal fusion for multimedia analysis: A survey," *J. Multimedia systems*, 2010.

[20] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2000.