Identifying Players in Broadcast Sports Videos using Conditional Random Fields

Wei-Lwun Lu, Jo-Anne Ting, Kevin P. Murphy, and James J. Little
University of British Columbia, Vancouver, BC, Canada
{vailen,jating,murphyk,little}@cs.ubc.ca

Abstract

We are interested in the problem of automatic tracking and identification of players in broadcast sport videos shot with a moving camera from a medium distance. While there are many good tracking systems, there are fewer methods that can identify the tracked players. Player identification is challenging in such videos due to blurry facial features (due to fast camera motion and low-resolution) and rarely visible jersey numbers (which, when visible, are deformed due to player movements). We introduce a new system consisting of three components: a robust tracking system, a robust person identification system, and a conditional random field (CRF) model that can perform joint probabilistic inference about the player identities. The resulting system is able to achieve a player recognition accuracy up to 85% on unlabeled NBA basketball clips.

1. Introduction

Our work addresses the problem of automatic tracking and identification of players in broadcast sport videos filmed at a medium distance with a single moving, zooming camera, as Figure 1 shows. In this paper, we focus on basketball, but the technique is general and applicable to other team sports such as football, ice hockey, soccer, etc. Labeled tracks will allow for automatic generation of game/player statistics and automatic camera control to track a specific player (e.g., the 2010 World Cup's Star Camera).

The problem is a challenging one and particularly relevant to the computer vision community for several reasons: (1) Tracking in sports is hard. Players move rapidly and unpredictably to make themselves hard to track. Tracking pedestrians is easier since they have similar and simpler, continuous motion patterns. (2) Tracking in a moving, zooming camera is harder since background subtraction is difficult. Coupled with motion blur, frequent occlusions and exit/re-entrance of players, tracking becomes non-trivial. (3) Identification is challenging. Faces are blurry and low-resolution ($\sim 15 \times 15$ pixels), making it impossible for even humans to identify players only from faces. Players on a team have the same jersey color and many have the same hair/skin color. Player dimensions may help but are hard to estimate since computing homographies in a moving camera is non-trivial.

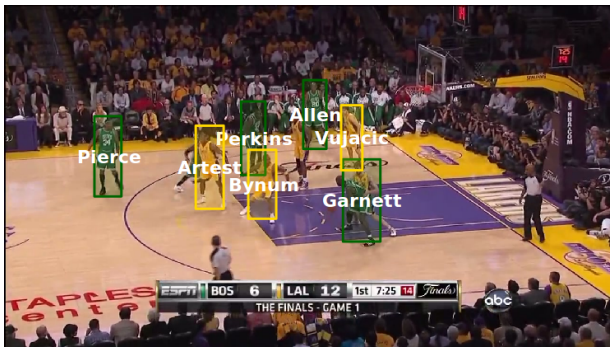
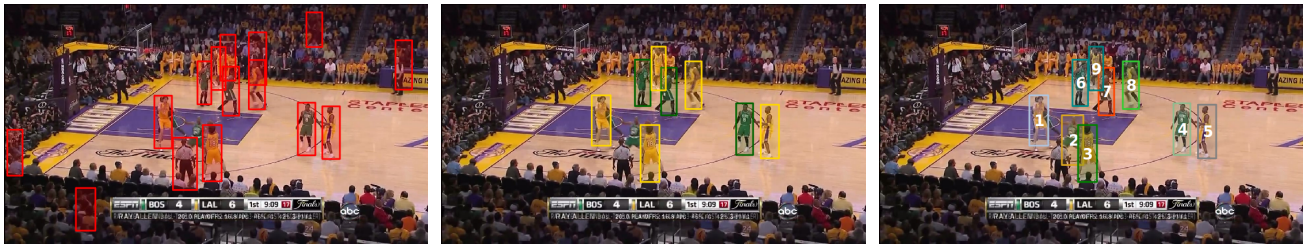


Figure 1: Automatic tracking and labeling of players from a medium-distance moving, zooming camera.

Most existing player identification systems focus on video clips taken from a close-up view where either facial features or numbers on the jersey are clear [1, 2, 3, 14, 25, 28]. In addition, most pedestrian re-identification techniques (e.g., [8, 12, 13]) cannot be applied directly since they typically rely heavily on each person having a unique color and/or shape. As far as we know, the problem of player tracking and identification in broadcast sports videos has yet to be solved. Our system is the first to do so.

The key idea of this paper is that we can propagate easy-to-classify images of a player to other images of the same player by using a conditional random field (CRF) [17] whose structure is created based on the output of a tracking system. The contributions of this paper is three-fold. First, we develop a tracking system that reliably tracks multiple players, even under severe occlusions. Second, we introduce a player appearance model that uses SIFT interest points [20], MSER regions [21] and color histograms. Third, we show how to perform joint classification of all



(a) Detection (b) Team classification (c) Tracking

Figure 2: (a) Automatic player detection generated by the DPM detector [9]. (b) Automatic team classification. Celtics are in green, and Lakers are in yellow. (c) Automatic tracking by associating detections into tracklets. Numbers represent the track ID not player identities.

player images by using a CRF. The structure of the CRF is built on top of the output of a tracking system, which links together images of the same player. In addition, we add mutual exclusion edges between all player images in a frame, which enforces the fact that a player can only appear once per frame. The CRF allows us to achieve player identification accuracies up to 85% in unlabeled test videos.

2. Related work

Reviewing all relevant tracking papers is beyond the scope of this paper, and we discuss only some of the most closely related work. [29] is a general survey of tracking systems. One key trend is the use of discriminative object detectors to help the generative tracking model. For example, Okuma *et al*. [24] used a Boosted Particle Filter (BPF) for tracking hockey players. Cai *et al*. [5] improved BPF by using bi-partite matching to associate detections with targets. Some systems first detect players and then associate detections with tracklets. For instance, Liu *et al*. [19] used data-driven MCMC (DD-MCMC) to create tracklets from detections and applied this technique to track soccer players. Ge *et al*. [11] not only used DD-MCMC to create longer tracks from shorter tracklets, but also learned parameters of the tracking system in an unsupervised manner.

Previous player identification systems in the sports domain have focused on videos taken from a close-up camera, and they rely on recognizing frontal faces, or numbers on the jersey. For instance, Bertini *et al*. [2, 3] trained face and number recognition systems using hand-labeled images and used the learned models to identify players on test videos. The system developed by Ballan *et al*. [1] used face matching. In order to improve matching accuracy under transformations, they extracted SIFT features [20] and performed a robust matching between faces. Ye *et al*. [28] relied on jersey number recognition, introducing an effective way to locate and segment the number region. Similarly, Saric *et al*. [25] performed jersey number recognition, but exploited color-based segmentation to extract the number region. Recently, Jie *et al*. [14] developed a player recognition system that relies on face and upper body pose. Our system is significantly different from past work because we address the problem of player identification from videos where facial features are blurred and jersey numbers are rarely seen.

There is a related problem in surveillance called pedestrian re-identification, where the goal is to find a pedestrian of some appearance over a network of cameras (e.g., [8, 12, 13]). Most of these systems rely on color, shape or texture and cannot be directly applied to sport videos due to the uniformity of jersey colors in a team. Some systems even use geometric configuration of cameras to help re-identify pedestrians (e.g., [13]), which is also not applicable in our case because we only have a single moving camera.

3. Automatic player tracking

In order to identify players, we have to first locate and track players over time, i.e., do multi-target tracking. This paper takes a *tracking-by-detection* approach, similar to [5, 11, 19, 24]. Specifically, we first run an object detector to locate players in every frame of a sports video, then we associate detections over frames with tracklets (a tracklet is a sequence of bounding boxes containing the same player over a period of time).

3.1. Player detection

We use the Deformable Part Model (DPM) detector developed by Felzenszwalb *et al*. [9] to automatically locate sport players. To train the DPM detector for basketball players, we first prepared 5000 positive patches of basketball players and 300 negative images containing no players. Then, we trained a DPM detector that consists of six parts and has two different aspect ratios for bounding boxes.

The DPM detector has a 69% precision and 73% recall. Figure 2(a) shows some DPM detection results in a sample basketball video. We observe that most false positives are generated from the spectators and referees, who have similar shapes to basketball players. Moreover, since the DPM detector applies non-maximum suppression after detection,

it may fail to detect players when they are partially occluded by other players.

3.2. Team classification

After player detection, we perform team classification to divide detected bounding boxes into three groups: Team A, Team B, and others. We do this for two reasons: (1) We want to ignore bounding boxes that correspond to spectators and referees since this is not the focus of the system. (2) Separating players into different teams simplifies the tracking and player identification problem because the number of targets is reduced by half.

Since players of a team wear uniforms of the same color, we use RGB color histograms as features to classify detected bounding boxes into one of three groups. Specifically, we first collect 1000 patches of Team A, Team B, and others (which include spectators, referees, and false positives). For every bounding box, we compute a 10-bin color histogram for each of the three RGB channels (resulting in a 30-bin color histogram). Since patches may contain background, we put a Gaussian weighting function centered in the patch to emphasize the central region. We then train a Logistic Regression classifier [4] with a L1 regularizer [23] for team classification. Figure 2(b) shows some team classification results in a basketball video.

3.3. Player tracking

Once detected players have been separated into teams, the next step is to associate detected bounding boxes over time with tracklets. Here, we take a *tracking-by-detection* approach, where the inputs are detections and outputs are tracklets of players.

We take a one-pass approach for tracking, similar to [5]. At any frame, we first associate detections with existing tracklets. To ensure a one-to-one matching between detections and tracklets, we perform bi-partite matching as in [5]. The matching scores are Euclidean distances between centers of bounding boxes and the predicted locations of players. We intentionally do not use colors in the matching score since players of a team wear the same uniform.

After assigning detections to existing tracklets, the next step is to update the state estimate of players. The state vector we want to track at time t is a 4-dimensional vector $\mathbf{x}_t = [x, y, w, h]^T$, where (x, y) represents the center of the bounding box, and (w, h) are its width and height, respectively. Let $\mathbf{z}_t = [x, y, w, h]^T$ be the detected bounding box at time t. We assume the following linear-Gaussian observation model: $p(\mathbf{z}_t|\mathbf{x}_t) = \mathcal{N}(\mathbf{z}_t|\mathbf{x}_t, \mathbf{\Sigma}_z)$, where $\mathbf{\Sigma}_z$ is a diagonal matrix set by hand.

Most tracking systems assume a first-order or second-order auto-regressive model for dynamics (e.g., [5, 24]). That is, they assume that $p(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t|\mathbf{A}\mathbf{x}_{t-1}, \mathbf{\Sigma}_x)$. More sophisticated models use Gaussian process regression to model the dynamics [27]. We got much better results by using a simpler model of the form $p(\mathbf{x}_t|\mathbf{x}_{t-1}, t) = \mathcal{N}(\mathbf{x}_t|t\mathbf{a}_t + \mathbf{b}_t, \mathbf{\Sigma}_x)$, where t is the current frame, $\mathbf{a}_t$ is a regression weight vector, and $\mathbf{b}_t$ is a regression offset term. The regression parameters $(\mathbf{a}_t, \mathbf{b}_t)$ are learned online based on a sliding window of data of the form $(t_i, \hat{\mathbf{x}}_{t_i})$, for $t_i = t - F, \ldots, t - 1$, where $\hat{\mathbf{x}}_{t_i}$ is the posterior mean state estimate at time t_i.

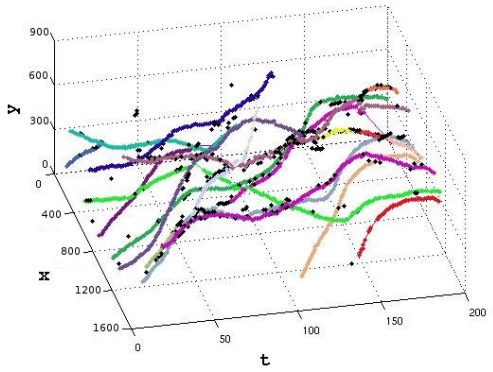


Figure 3: The x-y-t graph of tracking results, where (x, y) is the center of a bounding box, and t is the time. Every dot in the graph represents a detected bounding box, where different colors represent different tracklets.

Note that our motion model is independent of the previous state. However, it depends on the current time index. The reason it works well is that there is a local linear relationship between time t and the current state $\mathbf{x}_t$, as illustrated in Figure 3. Given our linear-Gaussian observation and motion models, we then update the current state using a Kalman Filter (KF) [16].

We create a new tracklet for detections that are not associated with any existing tracklets. This new tracklet will be first marked as *unreliable* until it has a certain number of detections associated with it. Otherwise, the new tracklet will be automatically dropped.

For tracklets that are not associated with any detection, the system will update the state using the prediction. However, if the tracklet does not have a detection over some time period (currently, 1 sec in experiments), it will be removed from the pool. The system will also terminate a tracklet when its bounding box moves out of the image border.

3.4. Results

To evaluate the quality of our tracker, we used Game 1 of the 2010 NBA Finals (Los Angeles Lakers vs. Boston Celtics) for testing. The video consists of different kinds of shots: close-up shots, medium-distance shots, and commercials. In this paper, we extracted 21 video clips from medium-distance shots, where the average length of video

(a) Raw image (b) MSER visual words (c) SIFT visual words

Figure 4: (a) Raw image patches extracted from the tracking system. (b) Green ellipses represent detected MSER regions [21]. (c) Red circles represent detected SIFT interest points [20].

clips is 796 frames, or 24.1 seconds.

Compared with ground-truth bounding boxes, the DPM detector has a a 69% precision and 73% recall. The team classifier then preserves those bounding boxes generated from players of both teams, while dropping those corresponding to spectators, referees and false positives. After team classification, we significantly increase the precision to 96% while sacrificing recall (which drops to 61%).

The tracking system comes in to create tracklets from detections. Since our tracking system predicts the player's location even when there is no associated detection, it is able to bridge the gap between temporally sparse detections. In this way, the tracking system retains a precision of 98%, while improving the recall to 81%, compared with ground-truth bounding boxes.

Figure 3 shows results of tracking basketball players in a test clip (see the online supplementary material for videos). Every dot in the graph represents the center of a detected bounding box, where different colors represent different tracklets. We can see that the proposed system is able to track multiple targets over long time intervals, even when players cross over each other. The motion model is able to predict a player's location even where there is no detection, e.g., tracklet #2 in Figure 2(c).

4. Learning player identities

Given the tracklets, our next step is to automatically identify the player each tracklet represents. Approaches relying on facial recognition are infeasible in our domain due to the far-field nature of most shots. For example, the average size of a player's head is about 15×15 pixels, and facial features are usually blurry due to a fast camera/player motion. It is very challenging even for human to identify players only from faces.

Recognizing numbers is possible, but still difficult. The numbers on the back of the jersey are 18×18, and numbers on the front of the jersey are even smaller at 10×10 pixels. We could train a number detector to locate candidate number regions before using an optical character recognition (OCR) module or even deep generative architectures like the convolutional neural net [18] to recognize jersey numbers. Disadvantages of the number recognition approach are: i) automatic number detection is hard in videos consisting of far-field shots; ii) gathering representative examples for training is labor-intensive because images of numbers are infrequent. We tried an off-the-shelf OCR system. However, it performed poorly on this data (results not shown) because of distorted numbers, shirt deformations, non-frontal viewpoints, etc.

We adopt a different approach, ignoring number recognition and focusing on identification of players as entities. Its main advantage is that we only need to label tracklets with their player class labels, which is easier than segmenting out number regions and labeling each digit. Below, we describe the features and classifier used and report results.

4.1. Features

We used a combination of three different features to generate visual words: maximally stable extremal regions (MSER) [10], SIFT visual words [20], and RGB color histograms. Visual words have been previously applied to object categorization (e.g., [26]).

- MSER regions [21] are stable segments whose colors are either darker or lighter than their surroundings. They are useful for detecting text in natural scenes because text has often uniform color and high contrast. In order to use MSER for player identification, we first detected MSER regions [21], as shown in Figure 4(b), and then normalized them [10]. For every MSER region, a 128-dimensional SIFT descriptor was computed and quantized into one of 300 visual words using a learned codebook (the codebook is learned using k-means clustering). The MSER representation of the image is a 300-dimensional bag-of-words bit vector, where a value of 1 indicates presence of the corresponding visual word in the image and 0 otherwise.

- SIFT interest points [20] are stable local patches that are invariant to scale and affine transformation. We first detected SIFT interest points, shown in Figure 4(c), and then extracted SIFT descriptors. The SIFT descriptors were quantized into 500 visual words (we

used more visual words for SIFT because there were more SIFT interest points than MSER regions).

- Although colors are weaker features (players of the same team wear the same uniform), skin color may provide some information for player identification. To account for colors of limbs, hair, etc., we also extracted RGB color histograms from the image. For the RGB color histogram, we used 10 bins for each of the R, G and B channels. We treat the three colors independently, so the full histogram has 30 bins/dimensions.

Figure 4 shows an example of MSER regions and SIFT interest points. We see that faces are always blurred, while numbers can only be seen clearly in the last frame. Since we do not segment the player from the background, some MSER regions and SIFT points are generated from the background, making player identification more challenging. The final feature vector for each image consists of 830 dimensions, where the first 800 dimensions are binary and the last 30 dimensions are positive values. Next, we discuss the effect of feature choice on player identification.

4.2. Classifiers

Using a combination of MSER + SIFT + color features, we train a classifier that maps image feature vectors to player class labels. We tried a mixture of classifier experts model [15], where each expert is an L1-regularized logistic regression [23]. We used a mixture of experts model because we hypothesized that each mixture component could learn a different view of the player (frontal, profile, etc.). However, using a single mixture component (i.e., a vanilla L1-regularized logistic regression model) worked just as well as using a mixture of 2 or 3 mixture components (results not shown). Since the vanilla model is faster to train (EM [7] is not necessary), we decided to use it for all subsequent experiments.

4.3. Results

To understand the effect of features on player identification, we trained a L1-regularized logistic regression model using various combinations of feature types. We used 12 video clips for training (a total of 9800 frames) and tested on 9 video clips[1]. Our evaluated dataset contained more than 15000 frames, which is similar in scale to benchmark datasets for moving cameras with annotated tracklets (e.g., ETHZ pedestrian dataset). Since tracklets are already classified into teams (done in automatic player tracking), we report classification accuracies on a per team basis. Although each team has 12 players (NBA rules), not all get to play.

[1]The shortest test clip was around 300 frames while the longest test clip had 1300 frames.

	Lakers		Celtics	
	I.I.D.	GM	I.I.D.	GM
MSER	33.67%	64.89%	36.04%	66.18%
SIFT	40.71%	67.56%	48.03%	73.89%
RGB	57.45%	69.70%	43.02%	48.02%
MSER + RGB	59.42%	75.10%	53.03%	70.60%
SIFT + RGB	58.52%	81.29%	57.05%	74.53%
MSER + SIFT	45.64%	77.18%	52.04%	78.58%
All 3	**61.86**%	**84.89**%	**60.85**%	**81.77**%

Table 1: Player classification accuracies as a function of features used, averaged over all test clips. I.I.D. represents classification done on a frame-per-frame basis (i.e., ignoring temporal coherence between detections). GM represents the graphical model in Figure 5 (see Section 5).

We exploit this fact to reduce the state space to $C = 9$ player classes per team[2].

Table 1 shows the classification accuracies, averaged over all test clips, where classification was done on I.I.D. detections (i.e., temporal coherence between detections in a tracklet were ignored). The results for I.I.D. classification shows that best results are obtained using all three types of features (MSER, SIFT and color) together. This conclusion still holds when using a CRF, as we discuss next.

5. The full system

No matter how good our features are, most detections are fundamentally ambiguous in their class label. We now describe our CRF model for performing joint classification, which allows us to borrow statistical strength from the reliable classifications to help ambiguous classifications.

5.1. The conditional random field

Once an appearance model is learned over player classes, we use it to infer player identities on unlabeled test videos. For each test video, we perform automatic player tracking (described in Section 3) to get tracklets. Let us assume that there are T frames in the test clip and D_t detections in frame t. We then construct a CRF [17], as Figure 5 shows. $\mathbf{x}_{td}$ represents the feature vector for the observed detection d in frame t. y_{td} represents the unknown identity for detection d in frame t (with C possible values).

Detections that belong to the same tracklet are connected with temporal edges with the following potential:

$$\psi_{time}(y_{tj}, y_{t+1,k}) = \begin{cases} 1-\epsilon & \text{if } y_{tj} = y_{t+1,k} \\ \epsilon & \text{otherwise} \end{cases} \quad (1)$$

[2]The Celtics had 10 active players in the videos considered, but one player plays for less than a minute so we removed him from consideration. The Lakers had 12 players reduced to 9 for the same reason.

where ϵ is a fixed parameter reflecting the amount of linking errors in the tracker. Setting $\epsilon = 0$ forces the identity of all linked detections to be identical.

Since all detections in a frame must be uniquely identified (no player can exist twice), we also introduce edges between the y_{td} nodes in each frame to enforce mutual exclusion in identities, using the following potential:

$$\psi_{mutex}(y_{tj}, y_{tk}) = \begin{cases} 1 & \text{if } y_{tj} \neq y_{tk} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

The overall model then takes the following form:

$$\begin{aligned} \log p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) \propto & \sum_{t=1}^{T} \sum_{d=1}^{D_t} \log p(\mathbf{y}_{td}|\mathbf{x}_{td}, \boldsymbol{\theta}) \\ & + \sum_{t=1}^{T} \sum_{d=1}^{D_t} \sum_{j=1, j\neq d}^{D_t} \log \psi_{mutex}(\mathbf{y}_{td}, \mathbf{y}_{tj}) \\ & + \sum_{t=1}^{T} \sum_{d=1}^{D_t} \sum_{j:succ(d,t)=j} \log \psi_{time}(\mathbf{y}_{td}, \mathbf{y}_{t+1,j}) \end{aligned} \quad (3)$$

where $succ(d, t)$ is the next node (if it exists) connected to y_{td} in the tracklet. Local evidence terms $p(y_{td}|\mathbf{x}_{td}, \boldsymbol{\theta})$ are computed by the logistic regression classifier.

5.2. Inference

For $C = 9$ and $D_t = 5$ detections, exact inference in the CRF in Figure 5 is intractable for long videos. We could perform exact inference in the model by merging all the nodes in each time slice and then treating the model as a form of a HMM. The state space of the collapsed model will be large, but exact inference is tractable[3]. However, exact inference in the HMM will not scale to other sports with larger teams (e.g., $C = 22$ for NHL hockey), and an approximate inference will be needed. For this reason, we focus on the CRF instead of the HMM variant.

We perform player identification separately for each team. We explored the following three variations on the full graphical model: i) an I.I.D. model (no edges between the y nodes, equivalent to treating each detection independently); ii) a graphical model with only mutex edges (edges between y nodes in a frame); iii) a graphical model with only temporal edges (edges between y nodes across frames).

For each of these 3 variants of Figure 5, we used exact inference, specifically the junction tree algorithm [4]. In the case of the model with only temporal edges, this is equivalent to running the forwards-backwards algorithm on each chain separately. In the case of the model with only mutex edges, exact inference is equivalent to exhaustive enumeration over all C^{D_t} configurations because all nodes are fully connected to each other.

[3]More precisely, the number of legal state configurations at frame t is $K_t = \sum_{d=0}^{D_t} C!/(C-d)!$. For $C = 12$ active players per team, and $D_t = 5$ detections per team per frame, we have $K_t = 108,385$.

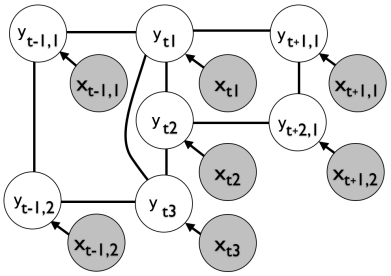


Figure 5: Graphical model of a test video: $\mathbf{x}$ are observed detections and $\mathbf{y}$ are unknown identities of detections. Mutex edges exist between $\mathbf{y}$ nodes within a frame. Temporal edges exist between $\mathbf{y}$ nodes across frames. Pairwise edges between $\mathbf{y}$ and $\mathbf{x}$ nodes exist in each frame.

For the full graphical model with mutex and temporal edges, junction tree is intractable due to large clique size. Instead, we used loopy belief propagation (BP) [22] to approximate inference, sacrificing accuracy for tractability. To give an idea of the speed, running loopy BP on the full graphical model for a video clip with 1090 frames and 4109 detections took 896 sec (Matlab code).

5.3. Results

Table 2 shows the player classification accuracies for both teams, averaged over all test clips for the the 4 graphical model variations. Results are for $\epsilon = 0.001$; results do not change for values of $0 < \epsilon \leq 0.01$ (equivalent to 1% error in tracking). Results for $\epsilon = 0$ are even worse for the full graphical model since automatically generated tracklets in the test videos have a small amount of tracking error.

We can draw two main conclusions from Table 2. First, adding temporal edges helps reduce the error the most. Adding mutex edges to a model with temporal edges also helps, despite the need to use approximate inference. The full CRF leads to a significant performance boost of up to 85% accuracy (from 62% accuracy in I.I.D. classification). Second, a sparse linear classifier performs very well, suggesting that the use of all three feature types (MSER, SIFT and RGB) is already quite discriminative (the "GM" columns of Table 1 confirm this as well).

Interestingly, test accuracies for the Celtics are slightly lower than that for the Lakers (82 vs. 85%). This can be explained by looking at the roster of active players. Of the 9 players on the Lakers, 3 are Caucasian: use of color as a feature allows for better discrimination of these 3 players. The Celtics, however, do not have this advantage, making

	Lakers	Celtics
I.I.D. model (no edges)	$61.86 \pm 2.64\%$	$60.85 \pm 8.16\%$
Graphical model with mutex edges only	$65.23 \pm 3.59\%$	$63.86 \pm 8.83\%$
Graphical model with temporal edges only	$78.25 \pm 6.95\%$	$78.92 \pm 7.73\%$
Full graphical model (loopy BP)	$\mathbf{84.89 \pm 6.06\%}$	$\mathbf{81.77 \pm 8.07\%}$

Table 2: Average player classification accuracies using L1-regularized logistic regression, with standard deviations shown. Results are averaged over all test video clips and reported for various configurations of the graphical model in Figure 5.

the identification problem harder.

Figure 6 shows qualitative tracking and identification results of the full system (videos are available online). Team classification results are shown with players from the Celtics in green bounding boxes and players from the Lakers in yellow boxes. Text within a bounding box indicates the player name predicted by the system. Bounding boxes in red represent misclassifications. We see that the system is able automatically track and identify players with relatively high accuracy.

We made several discoveries in the process of developing the system. For example, we found that jersey number recognition alone (with a number detector and OCR module) gave poor results. This is due to the fact that jersey numbers are infrequent and, when visible, are often deformed due to player movements. A second observation was that discriminative player models were more effective at weighting distinguishing features than generative ones, which can be attributed to the common feature points shared across players. We were also surprised to find that mutex constraints were not as effective as temporal constraints, which can be explained by frequent player occlusions. Our work sheds insights on how to solve the multi-player tracking and identification problem. It is important to note that even though identification relies on good tracking results, tracking errors and identity switches in tracklets would still be present due to frequent occlusions and failed detections. Identification, there, remains challenging.

6. Conclusions and future work

We address the challenging problem of automatic tracking and identification of players from broadcast sports videos shot from a medium-distance. We have shown how to develop a system that can achieve up to 85% accuracy in player identification in challenging basketball videos. Our current system relies on manually labeled data for training the classifier. We are currently working on including weak labels such as play-by-play text during training (e.g., similar to [6]). Future work will include making the model more robust to tracking errors and using semi-supervised learning to reduce the number of labels needed for training. With these improvements, we hope to be able to apply the system on a wide variety of sports videos, such as ice hockey, soccer, etc, with minimal human effort.

Acknowledgment

This work has been supported by grants from the Natural Sciences and Engineering Research Council of Canada, the Canadian Institute for Advanced Research, and the GEOIDE Network of Centres of Excellence. We thank Kenji Okuma, David Lowe, and Nando de Freitas for helpful comments and suggestions.

References

[1] L. Ballan, M. Bertini, A. D. Bimbo, and W. Nunziati. Soccer Players Identification based on Visual Local Features. In *CIVR*, 2007.

[2] M. Bertini, A. D. Bimbo, and W. Nunziati. Player Identification in Soccer Videos. In *MIR*, 2005.

[3] M. Bertini, A. D. Bimbo, and W. Nunziati. Automatic Detection of Player's Identity in Soccer Videos using Faces and Text Cues. In *ACM Multimedia*, 2006.

[4] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[5] Y. Cai, N. de Freitas, and J. J. Little. Robust Visual Tracking for Multiple Targets. In *ECCV*, 2006.

[6] T. Cour, B. Sapp, C. Jordan, and B. Taskar. Learning from Ambiguously Labeled Images. In *CVPR*, 2009.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

[8] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person Re-identification by Symmetry-Driven Accumulation of Local Features. In *CVPR*, 2010.

[9] P. Felzenszwalb, D. McAllester, and D. Ramanan. A Discriminatively Trained, Multiscale, Deformable Part Model. In *CVPR*, 2008.

[10] P.-E. Forssen and D. G. Lowe. Shape Descriptors for Maximally Stable Extremal Regions. In *ICCV*, 2007.

[11] W. Ge and R. T. Collins. Multi-target Data Association by Tracklets with Unsupervised Parameter Estimation. In *BMVC*, 2008.

[12] D. Gray and H. Tao. Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features. In *ECCV*, 2008.

[13] O. Javed, K. Shafique, Z. Rasheed, and M. Shah. Modeling Inter-camera Space-time and Appearance Relationships for

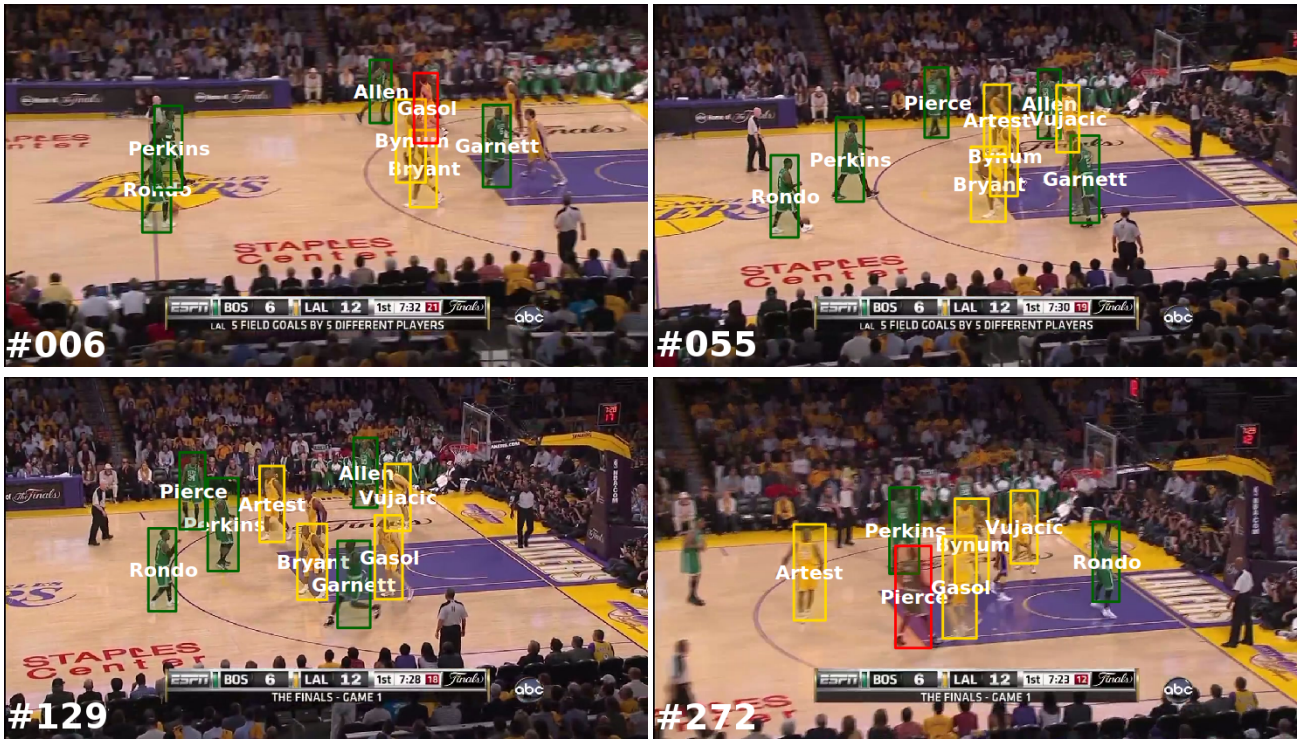


Figure 6: Automatic tracking and identification results in a broadcast basketball video. Green boxes represent Celtics players, and yellow boxes represent Lakers players. Text in boxes are automatic identification results (player's name), while red boxes highlight misclassifications. Frame numbers are on the bottom left corner.

Tracking across Non-overlapping Views. *Computer Vision and Image Understanding*, 109:146–162, 2008.

[14] L. Jie, B. Caputo, and V. Ferrari. Who's Doing What: Joint Modeling of Names and Verbs for Simultaneous Face and Pose Annotation. In *NIPS*, 2009.

[15] M. I. Jordan and R. A. Jacobs. Hierarchical Mixtures of Experts and the EM algorithm. *Neural Computation*, 6:181–213, 1994.

[16] R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45, 1960.

[17] J. Lafferty, A. McCallum, and F. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *ICML*, 2001.

[18] Y. LeCun, L. Bouttou, Y. Bengio, and P. Haffner. Gradient-based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[19] J. Liu, X. Tong, W. Li, T. Wang, Y. Zhang, and H. Wang. Automatic Player Detection, Labeling and Tracking in Broadcast Soccer Video. *Pattern Recognition Letters*, 30:103–113, 2009.

[20] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[21] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust Wide Baseline Stereo from Maximally Stable Extremal Regions. In *BMVC*, 2002.

[22] K. P. Murphy, Y. Weiss, and M. I. Jordan. Loopy Belief Propagation for Approximate Inference: An Empirical Study. In *UAI*, 1999.

[23] A. Ng. Feature Selection, L1 vs. L2 Regularization, and Rotational Invariance. In *ICML*, 2004.

[24] K. Okuma, A. Taleghani, N. de Freitas, J. J. Little, and D. G. Lowe. A Boosted Particle Filter: Multitarget Detection and Tracking. In *ECCV*, 2004.

[25] M. Saric, H. Dujmic, V. Papic, and N. Rozic. Player Number Localization and Recognition in Soccer Video using HSV Color Space and Internal Contours. In *ICSIP*, 2008.

[26] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman. Discovering Objects and Their Locations in Images. In *ICCV*, volume 1, pages 370–377, 2005.

[27] R. Urtasun, D. Fleet, and P. Fua. 3D People Tracking with Gaussian Process Dynamical Models. In *CVPR*, 2006.

[28] Q. Ye, Q. Huang, S. Jiang, Y. Liu, and W. Gao. Jersey Number Detection in Sports Video for Athlete Identification. In *SPIE*, 2005.

[29] A. Yilmaz and O. Javed. Object Tracking: A Survey. *ACM Computing Surveys*, 38(4):No. 13, 2006.