Finding People in Archive Films through Tracking

Xiaofeng Ren
Toyota Technological Institute at Chicago
1427 E. 60th Street, Chicago, IL 60637
xren@tti-c.org

Abstract

The goal of this work is to find all people in archive films. Challenges include low image quality, motion blur, partial occlusion, non-standard poses and crowded scenes. We base our approach on face detection and take a tracking/temporal approach to detection. Our tracker operates in two modes, following face detections whenever possible, switching to low-level tracking if face detection fails. With temporal correspondences established by tracking, we formulate detection as an inference problem in one-dimensional chains/tracks. We use a conditional random field model to integrate information across frames and to re-score tentative detections in tracks. Quantitative evaluations on full-length films show that the CRF-based temporal detector greatly improves face detection, increasing precision for about 30% *(suppressing isolated false positives) and at the same time boosting recall for over* 10% *(recovering difficult cases where face detectors fail).*

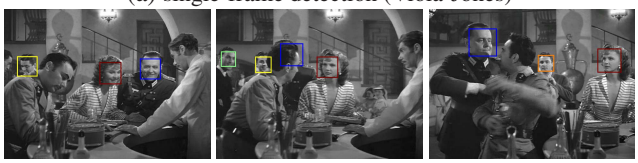
1. Introduction

Finding people is an important problem in computer vision and it offers an abundance of intellectual challenges, including, but not limited to, pose variation, self-occlusion and clothing. Archive films, such as *Casablanca* (1942), provides a particularly interesting setting: a quick look at the examples in Figure 1 would reveal several difficulties that are typical for these footages, such as low image quality, lack of color, frequent occlusion and crowded scenes.

Face detection is a natural starting point for analyzing these films: most of the time, faces of the actors and actresses are visible, either in a frontal or a profile view; only occasionally are they occluded or turn around and show the audience the back. In Figure 2(a) we show face detection results from the popular face detector of Viola and Jones [38]. Despite poor image quality, face detection performs reasonably well, finding most frontal faces under normal lighting conditions. Profile and rear views are problematic, however; and many false detections are present.

Figure 1. Archive films such as *Casablanca* (1942) offer an interesting and challenging setting for finding and tracking people. In this work we seek to find all people in such archive films, the groundtruth of which is visualized in these examples.

(a) single-frame detection (Viola-Jones)



(b) temporal detection through tracking

Figure 2. (a) Single-frame face detection (Viola-Jones) for three frames in *Casablanca*. It misses several people while introducing a fair number of false positives. (b) By tracking people through the scene, we integrate information temporally and achieve much better detection. Here we show top-ranked tracks indexed by color.

Temporal coherence comes to help where per-frame detection fails: real occurrences of people form consistent tracks in a video, while false positives tend to be isolated. If we can reliably track people through time, we can integrate information temporally and use it to find people missed by the single-image detector and to suppress false positives, as illustrated in Figure 2(b).

Tracking in these archive films, of course, is a highly

non-trivial problem itself. Low-level tracking is likely to fail because of the large variations in scale and appearance as well as partial occlusions. We take a detection-based strategy: whenever face detection is possible, we follow detections, hence avoiding the problem of modeling and updating appearance; we only switch to low-level tracking in cases where face detection is near-impossible, such as when a person turns around 180 degrees.

Once tracking has established temporal correspondences, detection becomes a one-dimensional inference problem on the tracks. We use a conditional random field (CRF) [23] to integrate information across frames. Quantitative evaluations on full-length films show that temporal integration greatly improves detection accuracy while at the same time finds people under difficult pose/illumination conditions. We also show the CRF-based integration outperforms several baselines including a local SVM classifier.

2. Related Work

Finding people in images is a broad field of research that can be roughly grouped into three areas: face detection, part-based detection, and whole-body template matching. Face detection is a well understood problem and efficient solutions have been found that perform really well under normal pose/illumination conditions [32, 38, 33]. Face detection under profile or non-standard views remains challenging [16, 40].

For body detection, when pose variation is limited (e.g. pedestrians) , template matching has been successful [14, 39, 11]. Part-based detection, explicitly modeling a person as an articulated object, can naturally tolerate pose variation [13, 19, 28] but becomes less robust. When only the upper body is visible, detection and pose estimation is difficult and largely unsolved [34, 24].

Tracking is a huge research field itself. The work of Lucas and Kanade [26] remains popular for low-level feature tracking. Linear systems and *Kalman Filtering* [15] are standard techniques to model dynamics. *Particle Filtering* has been shown to handle well non-Gaussian and multimodal distributions [20]. Region tracking can be made robust when there is enough information in color appearance [10]. Online appearance modeling has been shown to be useful [21, 9].

If we know we are tracking a person, there are several sources of knowledge that can be incorporated. Face tracking is popular and well explored (e.g. [17, 41, 25]). A lot of works in face tracking focus on facial features and expressions (e.g. [6]). Head tracking has been shown to be robust to changes in head pose and illumination [4]. On the other hand, tracking can be combined with a body model, either template-based [18, 5, 36, 29, 42] or part-based [7, 31]. Body tracking remains difficult under partial occlusions.

With efficient face detectors available, recently there has been an increasing interest in studying faces in large datasets or video (e.g. [3, 2, 12, 35, 30]). Most of these works focus on face identification rather than detection. Many take a clustering view and avoid tracking.

The work of Choudhury, Mikolajczyk and Schmid [8] presents an interesting and most relevant perspective. They use a particle filter to accumulate face detection probabilities temporally, and show that it improves detection performance. They use scores from a face detector only and there is no low-level appearance-based tracking. In particular, their approach only tolerates very short gaps between detections. Finding people in full-length archive films is more challenging and calls for a closer collaboration between detection and tracking.

3. Tracking People using Detection

Our tracking approach combines face detection and low-level tracking. The tracker operates in two modes: either a *detection step*, or a *track step*. Whenever the tracker can find and match a face detection, it follows the detection. This makes tracking robust to variations in scale, appearance and pose. If no face detection is available, the tracker has to rely on pixel appearance and continues tracking at low-level.

Low-level tracking is a well studied problem and many approaches have been proposed to reliably track objects (e.g. [20, 10, 21]). Our goal here is slightly different: we are not particularly interested in long-range tracking at low-level; instead our goal is to use low-level tracking to fill in gaps between face detections. Most of the time the gaps are short and tracking is easy. Only rarely people turn around or are occluded completely; in such cases tracking is very hard and it is unclear how much it affects the final detection performance. Based on these considerations, we choose to use methods that are simple and efficient: we use normalized correlation to track appearance, and use linear dynamics models and *Kalman Filtering* for smoothing.

3.1. Single-frame Face Detection

The initial step of our approach is to apply to each individual frame the Viola-Jones cascade face detector [38], as implemented in the Intel OPENCV library [1]. Both frontal and profile detectors are used. To compensate for poor image quality, we keep all the face candidates that pass into the last cascade (as suggested in [38]), effectively setting a very low threshold.

For each detection F, the real-valued output of the last cascade is an indicator of the "saliency" of the detection. We convert the output f into a log-likelihood score:

$$L(f) = \log\left[P_f(\text{face}|f)/(1 - P_f(\text{face}|f))\right]$$

where the empirical distribution P_f is estimated from data (being approximately Gaussian). We set a minimum of -3.

3.2. A Linear System for Tracking

We use a first-order linear dynamics model for position, and a zeroth order model for scale. Let $\{\mathbf{x}_t, \mathbf{v}_t, s_t\}$ be the position, velocity and scale of a face being tracked, we have

$$\begin{aligned} \mathbf{x}_{t+1} &= \mathbf{x}_t \quad +\mathbf{v}_t+ \quad \mathbf{N}(0; Q_{\mathbf{x}}) \\ \mathbf{v}_{t+1} &= \mathbf{v}_t \quad + \quad \mathbf{N}(0; Q_{\mathbf{v}}) \\ s_{t+1} &= s_t \quad + \quad \mathbf{N}(0; Q_s) \end{aligned}$$

We assume that $\tilde{\mathbf{x}}_t$ and $\tilde{s}_t$ are the observations of $\mathbf{x}_t$ and s_t, with additive Gaussian noises of covariance $R_{\mathbf{x}}$ and R_s.

The classical *Kalman Filter* [15] provides an optimal estimate of the linear system. For each track, at time t we have model predictions $\left\{\hat{\mathbf{x}}_{t|t-1}, \hat{s}_{t|t-1}\right\}$, based on observations from time $1, \cdots, t-1$. We then search for face detections around the position $\hat{\mathbf{x}}_{t|t-1}$ and scale $\hat{s}_{t|t-1}$.

For any face detection F found in the initial detection stage, let $\mathbf{x}_f$ be its position and s_f be its scale. we use a fixed threshold α_{match}: F is a match for this track if $\|\hat{\mathbf{x}}_{t|t-1} - \mathbf{x}_f\|_\infty < \alpha_{match}\hat{s}_{t|t-1}$ and $|\hat{s}_{t|t-1} - s_f| < \alpha_{match}\hat{s}_{t|t-1}$. α_{match} is set to 0.3 in our experiments.

At any time, there are potentially a large number of active tracks competing for face detections. Only a few are "good" tracks that have been reliably tracked; many are spurious ones. We would like the "good" tracks to have priority in matching detections. To capture this intuition, We use a greedy strategy for data association: we assign a track score to each active track, and let the tracks match and select face detections in the descending order of their scores. We score an active track by counting N_{detect}, the number of detections it has matched and accumulated.

3.3. Low-level Tracking with Correlation

In archive films, there are many situations in which a face detector could fail: the image quality may be poor; the lighting may be wrong, or the person may be facing away. If, at time t, an active track cannot find and match a detection, we switch to a low-level tracking mode, and use image pixels to continue tracking. We maintain a head template T_{head}, and use normalized correlation to search for the best appearance-based match in the image. Specifically, we search in a small window (10×20 pixels) around the predicted location $\hat{\mathbf{x}}_{t|t-1}$. The best matched location is the new observation $\tilde{\mathbf{x}}_t$.

After incorporating the observation in the current frame, we have the updated estimates $\hat{\mathbf{x}}_{t|t}$ and $\hat{s}_{t|t}$, based on all observations at time $1, \cdots, t$. Let $I(\hat{\mathbf{x}}_{t|t}, \hat{s}_{t|t})$ be the image patch at location $\hat{\mathbf{x}}_{t|t}$ and size $\hat{s}_{t|t}$. If t is a detection step, we set $T_{head} = I(\hat{\mathbf{x}}_{t|t}, \hat{s}_{t|t})$ (i.e. completely trusting the current detection). Otherwise, if t is a track step, we update the template linearly to be $T_{head} = (1 - \beta_{update})T_{head} + \beta_{update} I(\hat{\mathbf{x}}_{t|t}, \hat{s}_{t|t})$. β_{update} is set to be 0.1.

3.4. Initialization and Termination

Our initialization strategy is simple: whenever a face detection in the current frame cannot be matched to any of the active tracks, we use it to start a new track. A more conservative strategy would be to start a track only at highly reliably detections. It would be less prone to error but may miss many true faces and may require tracking both forward and backward. We keep tracking in the on-line fashion and rely on the greedy data association to tolerate spurious tracks.

To check for termination, we keep track of N_{detect}, the number of detection steps, and N_{track}, the number of correlation track steps. We terminate a track if $N_{track}/N_{detect} > 1.5$. Most spurious tracks are terminated in a few steps.

3.5. Extensions

To improve our simple correlation-based tracker, we use two extensions: first, we try to capture the intuition that, during a track step, a low correlation score indicates low confidence in the observation. Therefore, if t is a track step and the maximum correlation score is C_t, we add an additional term to the observation covariance $R_{\mathbf{x}}$, using $R_{\mathbf{x}}(t) = R_{\mathbf{x}} + R_{corr} * (1 - C_t)$. We set $R_{corr} = 100$.

In a second extension, we try to compensate for the drifting issue in low-level tracking [27]. Instead of one template for the head, we use two templates T^0_{head} and T^1_{head}. We update one of the templates, say T^1_{head}, as normal. The other template T^0_{head} is kept fixed during track steps. The correlation score C_{head} becomes a weighted average of two, $w^0 C^0_{head} + (1 - w^0)C^1_{head}$. We set w^0 to be 0.75.

4. Detection through Temporal Integration

In the previous section we have discussed our simple tracker combining per-frame face detection and correlation-based tracking. Tracking establishes temporal correspondences across frames, and we will now explore this correspondence to integrate information over time.

Previous approaches to face detection and tracking take a simplified view of temporal reasoning: either tracking is successful, and every face in the track is declared a positive; or it is found unsuccessful, and the whole track is discarded. They typically put a stringent condition on tracking to ensure no error in temporal correspondence. For example, the work of [8] allows a maximum of five track steps.

In our setting of archive films, it is fairly common for the face detector to miss a person for a long period of time. As we observe from the examples in Figure 3, low-level tracking finds temporal correspondences under many difficult conditions, and long-range correspondences can be successful. Of course, the longer we track without correction from a detection, the less reliable the correspondence becomes. Any model of temporal integration would have to take this uncertainty into account.

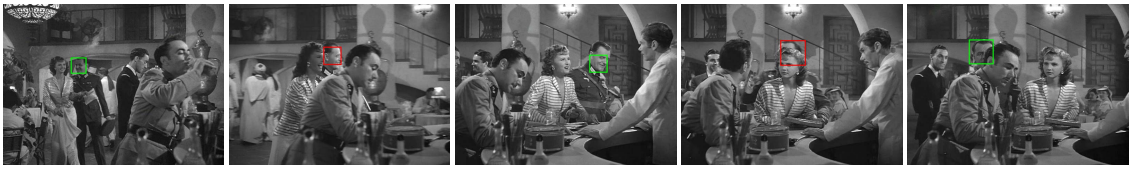


Figure 3. Tracking examples combining detection and correlation tracking. A detection step (when the track finds and matches a face detection close by) is shown in green, and a track step (using correlation) is shown in red. First row: #1, initial detection; #48, detection avoids drifting under poor image quality and background change; #203, the face, under a strange pose, is not detected, and the tracker switches to correlation tracking; #334, tracking continues when the head turns 180 degrees; #378, a detection corrects tracking and fixates it back on the person. Second row: a dynamic model helps tracking under occlusions. The track is lost after #48 and restarts at #81. After that, the track successfully goes through the shot, under severe occlusions such as in #347-#373.

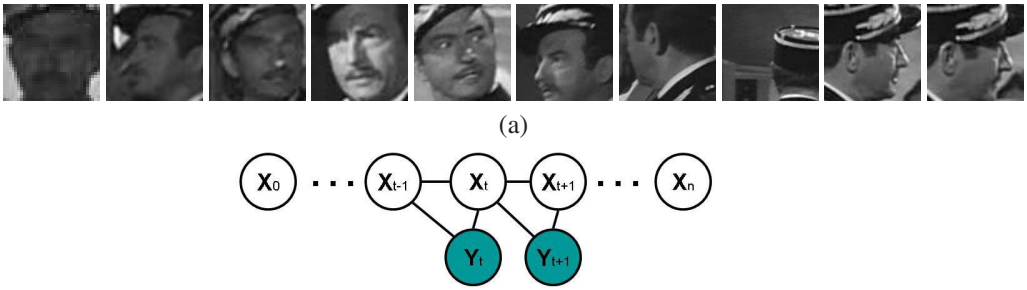


Figure 4. (a) Tracking establishes temporal correspondences between tentative faces, organizing them into linear chains. Temporal coherence can then be explored to improve detection and suppress noise. (b) We use a conditional random field (CRF) model for temporal integration. An observation $\mathbf{Y}_t$ affects both the likelihood for the binary label $\mathbf{X}_t$ and the coupling between adjacent labels $\mathbf{X}_{t-1}$ and $\mathbf{X}_t$. A CRF model provides a principled way to model the couplings and to integrate local evidences over time.

We formulate temporal integration as a probabilistic inference problem. For each step t in a track, we associate a binary random variable $\mathbf{X}_t$, $\mathbf{X}_t = 1$ if it predicts a person, and $\mathbf{X}_t = 0$ if not[1]. Tracking provides temporal correspondence that couples these variables in a one-dimensional track/chain. Linear-chain probabilistic models are well-understood and widely used. Many can readily apply here.

4.1. Conditional random field for temporal scoring

We develop a one-dimensional *conditional random field* (CRF) [23] to model temporal integration and to detect people in tracks. Consider a track of length n: let $\mathbf{X} = \{\mathbf{X}_t\}$ be the hidden variables, and let $\mathbf{Y} = \{\mathbf{Y}_t\}$ be the observations. We define the joint distribution of $\{\mathbf{X}_t\}$ as follows:

$$P(\mathbf{X}|\mathbf{Y}) = \frac{1}{\mathbf{Z}(\mathbf{Y};\Lambda)} \exp\left\{\sum \lambda_k f_k(\mathbf{X}_{t-1}, \mathbf{X}_t, \mathbf{Y}_t)\right\}$$

[1]Note that faces in a single track may take on different labels.

where $\Lambda = \{\lambda_k\}$ is a set of weights for the features $\{f_k\}$, and $\mathbf{Z}(\mathbf{X};\Lambda)$ is the normalization constant or the partition function (see Figure 4). Inference in this one-dimensional CRF model is solved by belief propagation.

Each observation $\mathbf{Y}_t$ has several components: δ_t, $= 1$ if t is a detection step, $= 0$ if a track step; L_t, the face saliency/likelihood score from Viola-Jones, defined when $\delta_t = 1$; and C_t, the correlation score, defined when $\delta_t = 0$.

We use two sets of features in the CRF model. One set of features relates the observation $\mathbf{Y}_t$ to the hidden variable $\mathbf{X}_t$: we use $f_1 = \mathbf{1}_{(\mathbf{X}_t=1)}$; $f_2 = \delta_t \mathbf{1}_{(\mathbf{X}_t=1)}$; and $f_3 = L_t \delta_t \mathbf{1}_{(\mathbf{X}_t=1)}$, in which we incorporate detection scores. We also add a prior on the first and last nodes of the track, $f_0 = \mathbf{1}_{(\mathbf{X}_0=1)} + \mathbf{1}_{(\mathbf{X}_n=1)}$.

The other set of features models the interactions or couplings between adjacent variables: we let $f_4 = \delta_t \mathbf{1}_{(\mathbf{X}_{t-1}=\mathbf{X}_t)}$, $f_5 = (1-\delta_t)\mathbf{1}_{(\mathbf{X}_{t-1}=\mathbf{X}_t)}$, and $f_6 = C_t(1-\delta_t)\mathbf{1}_{(\mathbf{X}_{t-1}=\mathbf{X}_t)}$. By introducing these features, we make the coupling between $\mathbf{X}_{t-1}$ and $\mathbf{X}_t$ dependent on the observa-

tions δ_t and C_t. Hence the model is capable of representing a stronger coupling during a detection step ($\delta_t = 1$) and a weaker coupling during a track step ($\delta_t = 0$). The coupling during a track step may also be dependent on the correlation score C_t; the lower C_t is, the less confident we are about the temporal coherence between $t-1$ and t.

Tracks co-exist in the same frames, so naturally there are interactions. One such interaction is *mutual exclusion*: unless being severely occluded, a person spans a certain zone of support, and it is unlikely to find another person within his zone. For example, in Figure 2(a), if we have detected the lady at the center, we would have a lower confidence for the detections on her dress.

It would be hard to model such inter-track relations exactly, and it would greatly increase the computational cost. Hence we use an approximation: we keep each track separate, and add an additional observation e_t for mutual exclusion. Let $\mathbf{x}_t^{(i)}$ be the center location of face t in track i. If there exists another track j such that $\mathbf{x}_t^{(i)}$ falls in the spatial support of $\mathbf{x}_{t'}^{(j)}$ (defined by two rectangles for head-body), and if j is a "better" track (here we approximate the score of a track with N_{detect}, the number of detections), we set the conflict variable $e_t^{(i)} = 1$, and introduce a new feature $f_7 = e_t \delta_t \mathbf{1}_{(\mathbf{x}_t=1)}$ in track i.

One nice property of the conditional random field model is that it comes with an elegant solution for parameter estimation, with the gradient of the log-likelihood being a difference between two expectations, one under the empirical distribution, and one under the model. Working with full-length films, we only have groundtruth labels on a sparse set of frames. Nevertheless, we can still compute the gradient and maximize the log-likelihood for the partial labeling. (A similar example of partial labeling can be found in [37]).

Let U be the set of all indices $\{1, \cdots, n\}$, S be the subset of U that we have groundtruth on, and $\bar{S} = U \backslash S$. Let $\mathbf{x}_S$ be the groundtruth labels for the variables $\mathbf{X}_S$. Omitting $\mathbf{Y}$ in the formulas for clarity, the likelihood $P(\mathbf{X}_S = \mathbf{x}_s)$ is the marginalization over the variables in $\bar{S}$:

$$P(\mathbf{x}_S) = \sum_{\mathbf{X}_t : t \in \bar{S}} \frac{1}{\mathbf{Z}(\Lambda)} e^{\sum \lambda_k f_k} = \frac{\sum_{\mathbf{X}_t : t \in \bar{S}} e^{\sum \lambda_k f_k}}{\sum_{\mathbf{X}_t : t \in U} e^{\sum \lambda_k f_k}}$$

The gradient of the log-likelihood $\mathbf{L}(\mathbf{x}_S) = \log P(\mathbf{x}_s)$ w.r.t to a parameter λ_q is

$$\begin{aligned} \frac{\partial}{\partial \lambda_q} \mathbf{L}(\mathbf{x}_S) &= \frac{\sum_{\mathbf{X}_t : t \in \bar{S}} f_q e^{\sum \lambda_k f_k}}{\sum_{\mathbf{X}_t : t \in \bar{S}} e^{\sum \lambda_k f_k}} - \frac{\sum_{\mathbf{X}_t : t \in U} f_q e^{\sum \lambda_k f_k}}{\sum_{\mathbf{X}_t : t \in U} e^{\sum \lambda_k f_k}} \\ &= \langle f_q \rangle |_{\mathbf{X}_S = \mathbf{x}_S} - \langle f_q \rangle \end{aligned}$$

where the second term is the expectation of f_q, and the first is the expectation of f_q conditioned on the observed labels $\mathbf{x}_S$. Both expectations can be computed using belief propagation. We use *stochastic gradient* to maximize $\mathbf{L}(\mathbf{x}_S)$.

4.2. Baseline models

In our experiments we compare our CRF temporal integration model to several baseline scoring schemes. One obvious choice of a baseline is N_{detect}, the number of detections accumulated in a track. This matches the intuition that the longer a track is (counting actual detections), the more likely it is a consistent track of people. We have used it as an approximate score in computing mutual exclusion. This score becomes less meaningful when comparing tracks from different shots, as the lengths of the shots vary greatly.

Another obvious choice is the average face score in a track, $\bar{L} = (\sum L_t \delta_t)/N_{detect}$. This will boost the score for weak detections in a track, by combining evidences from strong detections. This score ignores the length of the track.

One may also try putting these two together, by computing the sum of the face scores $L_{sum} = \sum L_t \delta_t$. This score favors both long tracks and tracks with strong face detections. On the other hand, it has a strong bias toward long tracks and ignores any of the track steps.

Finally, we can combine the features above in a local classification model using a *support vector machine* (SVM). The features N_{detect}, $\bar{L}$ and L_{sum} are all global features, i.e. defined on entire tracks; we use another global feature N_{track}, the number of track steps. We also include local features, defined at each t, including δ_t, L_t, e_t and C_t. These features are combined in a SVM with a linear kernel, using SVM^{light} [22].

5. Experiments

We conduct our experiments on three full-length black-white archive films: *Casablanca* (1942), 147600 frames of resolution 464×640; *Kind Hearts and Coronets* (1949), 153478 frames of resolution 448×655; and *The Great Dictator* (1940), 215405 frames of resolution 480×720. Human subjects mark groundtruth faces in a subset of the frames (every 50th in *Casablanca*, and every 100th in the other two films). We use the first half of *Casablanca* for training and use the rest for testing.

To avoid tracking across shots, we use two simple features to locate shot boundaries: normalized correlation between adjacent frames, and the derivative of the normalized correlation. *Casablanca* is automatically divided into 744 shots. The other two films are partitioned similarly.

The groundtruth faces are too sparse (in time) to be useful for training our tracking algorithm. Most of the parameters in our tracking model are covariances of the transition and observation models; standard linear system theory estimates them using *Expectation-Maximization*. We train the parameters as follows: first we start with a set of hand-set parameters. Given the initial tracking results, we use "good" tracks (tracks with length > 50) to estimate the system covariances Q. Finally, we use all the tracks to learn

the observation covariances R.

We use groundtruth labels to train our CRF model. We again use "good" tracks (length > 50) only in training. Stochastic gradient works well in our experiments, converging much faster than conjugate gradient. Despite the sparseness in labeling, the parameters learned are quite sensible: we find that coupling is stronger in a detection step, with $\lambda = 7.77$, comparing to $\lambda = 3.69$ in a track step; we also find a positive dependence between coupling and correlation score.

We evaluate the performance of a face/person detector by *Precision-Recall*. A detection matches to a groundtruth face if the overlapping area is at least 40% of both rectangles. We vary the threshold on the detector output, computing recall as the percentage of groundtruth people being found, and precision as the percentage of detections that are matched to groundtruth.

In Figure 5(a) we show the quantitative evaluation of our CRF-based temporal integration detector on *Casablanca*. We greatly improve the detection precision,from about 60% to 90%, while also improving recall from 58% to 70%. In Figure 5(b), we show that the CRF-based integration model indeed outperforms the baseline schemes, showing that temporal integration in tracks is a non-trivial problem.

In Figure 6, we show the experimental results on the other two films. We use the same tracking and integration parameters as trained from *Casablanca*. Again we see that temporal integration offers large improvements in both precision and recall. All the precision-recall curves look qualitatively the same.

In Table 1 we list the average precision (area under P/R curves), up to 70% recall, for all methods tested. The empirical results on the three films, hundreds of thousands of frames in each, are surprisingly similar: in all three cases, temporal integration improves average precision by about 30%, and improves asymptotic recall by about 12%.

The CRF-based scoring model assigns a score for each face/person in each track. We can compute a track score by averaging scores of the faces in it. In Figure 7 we show top-ranked tracks indexed by color. Comparing to single-frame detection (see the examples in Figure 2), we are able to find most people in the scenes and rank them above false detections, in challenging situations such as low image quality, motion blur, non-standard pose and partial occlusion. Complete tracks are available as supplementary materials.

6. Discussions

In this work we have studied the challenging problem of finding people in archive films. We take a temporal integration approach, exploring the synergies between face detection and tracking. Detection makes tracking robust, working under occlusion, pose/illumination variation and poor image quality. Tracking establishes temporal correspondences and a one-dimensional CRF model effectively integrates information along tracks, greatly improving both the precision and recall of face detection.

We have kept the components of our approach conceptually simple. There are a number of natural extensions, such as employing a sophisticated low-level tracking (e.g. combining region tracking with local features) or building better online appearances models.

The asymptotic recall of our detector is about 70% in all three films. As we can see in Figure 1, there are many people in these films that would be very difficult to locate. It remains to be seen how far we could push the recall rate without lowering precision much. One would probably need a more semantic understanding of the scenes.

Acknowledgments. We thank Deva Ramanan for many helpful discussions and suggestions.

References

[1] OpenCV, http://sourceforge.net/projects/opencvlibrary/.

[2] O. Arandjelovic and A. Zisserman. Automatic face recogntion for film character retrieval in feature-length films. In *CVPR*, volume 1, pages 860–7, 2005.

[3] T. Berg, A. Berg, M. Maire, R. White, Y. Teh, E. Miller, and D. Forsyth. Names and faces in the news. In *CVPR*, 2004.

[4] S. Birchfield. Elliptical head tracking using intensity gradients and color histograms. In *CVPR*, pages 232–7, 1998.

[5] M. Black and A. Jepson. Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. *Int'l. J. Comp. Vision*, 26(1):63–84, 1998.

[6] M. Black and Y. Yacoob. Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion. In *ICCV*, pages 374–381, 1995.

[7] C. Bregler and J. Malik. Tracking people with twists and exponential maps. In *CVPR*, pages 8–15, 1998.

[8] R. Choudhury, C. Schmid, and K. Mikolajczyk. Face detection and tracking in a video by propagating detection probabilities. *IEEE Trans. PAMI*, 25(10), 2003.

[9] R. Collins and Y. Liu. On-line selection of discriminative tracking features. In *ICCV*, pages I:346–352, 2003.

[10] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *CVPR*, pages II:142–149, 2000.

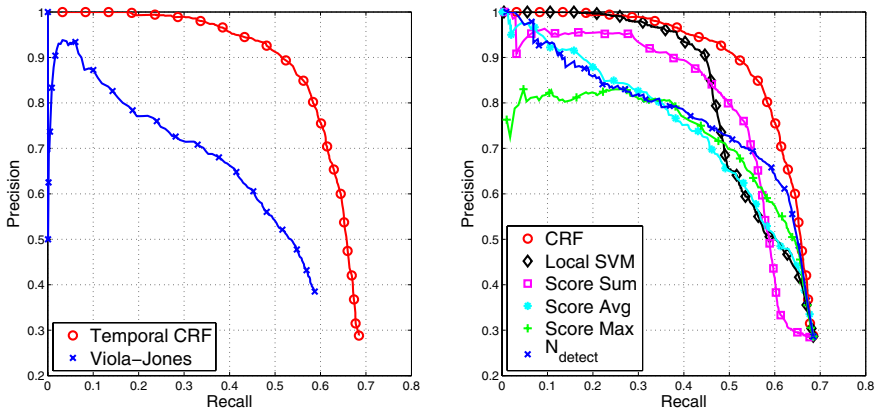
[11] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages I:886–893, 2005.

[12] M. Everingham and A. Zisserman. Identifying individuals in video by combining generative and discriminative head models. In *ICCV*, pages II:1103–1110, 2005.

[13] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *Int'l. J. Comp. Vision*, 61(1):55–79, 2005.

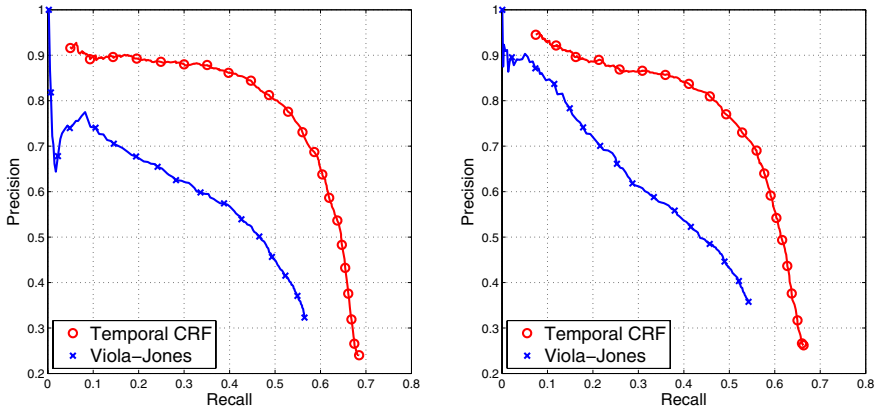
[14] D. Gavrila and V. Philomin. Real-time object detection for smart vehicles. In *ICCV*, pages 87–93, 1999.

[15] A. Gelb, editor. *Applied Optimal Estimation*. MIT Press, 1974.



(a) (b)

Figure 5. Quantitative evaluations on *Casablanca* (1942), using precision-recall: (a) we compare our CRF temporal integration approach to the Viola-Jones face detector (which we use as input). Temporal integration greatly improves the detection: for a wide range of recall ($0\% - 70\%$), the average precision increases from about 60% to 90%. We also increase the asymptotic recall from 58% to 70%, suggesting that the tracking algorithm finds a fair number of people where face detector fails (even with a low threshold). (b) We compare the CRF model to a few baselines, including a local SVM classifier using a combination of local and global features. The CRF model performs the best, especially at the high-recall range (i.e. the hard cases).



(a) *Kind Hearts and Coronets* (b) *The Great Dictator*

Figure 6. Further evaluation of tracking-based detection on two other films, *Kind Hearts and Coronets* (1949) and *The Great Dictator* (1940). We show precision-recall curves for both the single-image Viola-Jones detector and the CRF-based temporal detector. Just as in *Casablanca*, temporal integration greatly improves detection performance, about 30% increase in precision and 12% in recall.

[16] E. Hjelmas and B. K. Low. Face detection: a survey. *Computer Vision and Image Understanding*, 83:236–74, 2002.

[17] G. Hua and Y. Wu. Multi-scale visual tracking by sequential belief propagation. In *CVPR*, pages I:826–833, 2004.

[18] D. P. Huttenlocher, J. J. Noh, and W. J. Rucklidge. Tracking non-rigid objects in complex scenes. In *ICCV*, pages 93–101, 1993.

[19] S. Ioffe and D. Forsyth. Probabilistic methods for finding people. *Int'l. J. Comp. Vision*, 43(1):45–68, 2001.

[20] M. Isard and A. Blake. Condensation: Conditional density propagation for visual tracking. *Int'l. J. Comp. Vision*, 29(1):5–28, 1998.

[21] A. Jepson, D. Fleet, and T. El-Maraghi. Robust online appearance models for visual tracking. *IEEE Trans. PAMI*, 25(10):1296–1311, 2003.

[22] T. Joachims. Making large-scale svm learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Adv. in Kernel Methods - Support Vector Learning*. MIT-Press, 1999.

[23] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, pages 282–289, 2001.

[24] M. Lee and I. Cohen. Human upper body pose estimation in static images. In *ECCV*, pages 126–138, 2004.

[25] Y. Li, H. Ai, T. Yamashita, S. Lao, , and M. Kawade. Tracking in low frame rate video: a cascade particle filter with discriminative observers of different lifespans. In *CVPR*, pages I:1–8, 2007.

[26] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proc. of IJCAI*, pages 674–679, 1981.

	Viola-Jones	Temporal Integration					
		N_{detect}	Score-Max	Score-Avg	Score-Sum	Local SVM	**CRF**
Casablanca	**0.596**	0.768	0.715	0.740	0.791	0.818	**0.891**
Coronets	**0.495**	0.696	0.639	0.636	0.733	0.700	**0.805**
Dictator	**0.510**	0.596	0.588	0.618	0.653	0.642	**0.763**

Table 1. Average precision evaluation (up to 70% recall) on the three films. Temporal integration increases average precision by about 30%. The CRF-based integration performs the best, about 10% higher than the second-best strategy (Score-Sum).

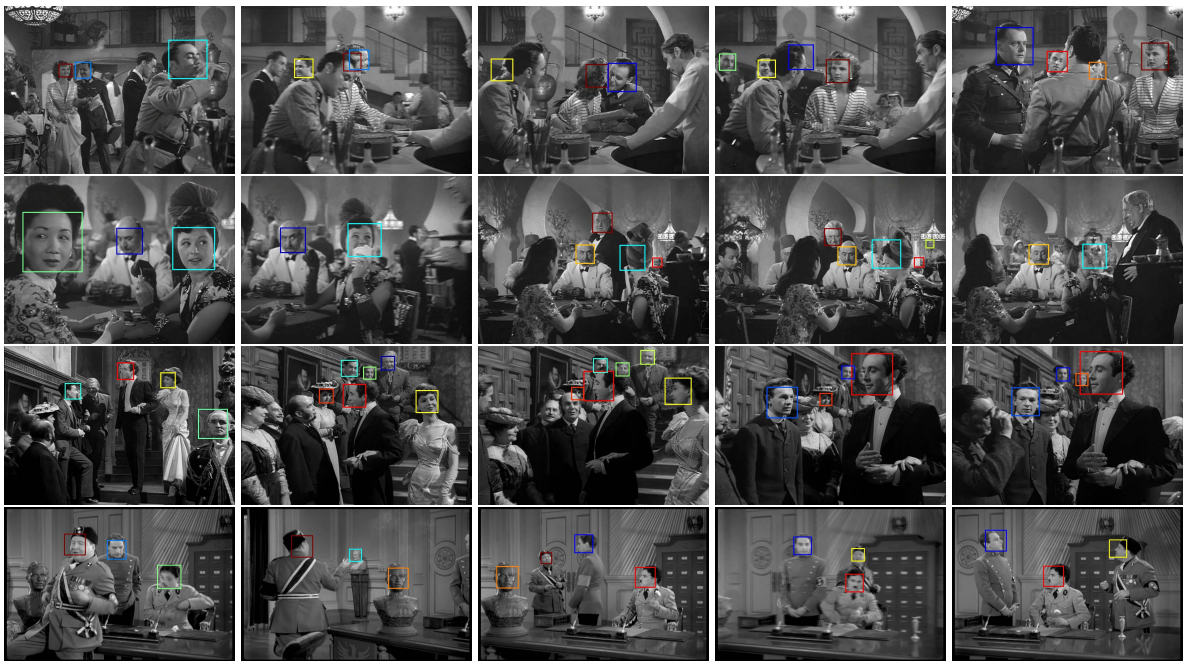


Figure 7. Examples of top-ranked tracks in the three archive films, indexed by color. Through temporal integration, we are able to find most people, with few false positives, in challenging situations such as large scale variation, partial occlusion, non-typical poses and crowded scenes. Complete tracks are available as supplementary materials.

[27] I. Matthews, T. Ishikawa, and S. Baker. The template update problem. In *BMVC*, pages II:649–658, 2003.

[28] G. Mori, X. Ren, A. Efros, and J. Malik. Recovering human body configurations: combining segmentation and recognition. In *CVPR*, volume 2, pages 326–333, 2004.

[29] K. Okuma, A. Taleghani, N. de Freitas, J. Little, and D. Lowe. A boosted particle filter: Multitarget detection and tracking. In *ECCV*, pages 28–39, 2004.

[30] D. Ramanan, S. Baker, and S. Kakade. Leveraging archival video for building face datasets. In *ICCV*, pages I:1–8, 2007.

[31] D. Ramanan, D. Forsyth, and A. Zisserman. Strike a pose: tracking people by finding stylized poses. In *CVPR*, 2005.

[32] H. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Trans. PAMI*, 20:23–28, 1998.

[33] H. Schneiderman and T. Kanade. A statistical method for 3d object detection applied to faces and cars. In *CVPR*, 2000.

[34] G. Shakhnarovich, P. Viola, and T. Darrel. Fast pose estimation with parameter sensitive hashing. In *ICCV*, pages II:750–757, 2003.

[35] J. Sivic, L. Ztinick, and R. Szeliski. Finding people in repeated shots of the same scene. In *BMVC*, pages III:909–918, 2006.

[36] K. Toyama and A. Blake. Probabilistic tracking in a metric space. In *ICCV*, pages II:50–57, 2001.

[37] J. Verbeek and B. Triggs. Scene segmentation with crfs learned from partially labeled images. In *Advances in Neural Information Processing Systems 20*, 2007.

[38] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, pages I:511–517, 2001.

[39] P. Viola, M. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. In *ICCV*, pages 734–741, 2003.

[40] P. Wang and Q. Ji. Learning discriminant features for multi-view face and eye detection. In *CVPR*, pages 373–9, 2005.

[41] T. Yang, S. Z. Li, Q. Pan, J. Li, and C. Zhao. Reliable and fast tracking of faces under varying pose. In *Conf. on Automatic Face and Gesture Recognition*, pages 421–8, 2006.

[42] T. Zhao and R. Nevatia. Tracking multiple humans in crowded environment. In *CVPR*, pages II:406–13, 2004.