Proceedings Paper:
Tavanai, A, Sridhar, M, Gu, F et al. (2 more authors) (2014) Context aware detection and tracking. In: Proceedings - International Conference on Pattern Recognition. 2014 22nd International Conference on Pattern Recognition (ICPR), 24-28 Aug 2014, Stockholm. IEEE , 2197 - 2202. ISBN 9781479952083

https://doi.org/10.1109/ICPR.2014.382

Context Aware Detection and Tracking

Aryana Tavanai, Muralikrishna Sridhar, Feng Gu, Anthony G. Cohn and David C. Hogg
School of Computing, University of Leeds
Leeds, LS29JT, United Kingdom
{fy06at,scms,f.gu,a.g.cohn,d.c.hogg}@leeds.ac.uk

***Abstract*—This paper presents a novel approach to incorporate multiple contextual factors into a tracking process, for the purpose of reducing false positive detections. While much previous work has focused on improving object detection on static images using context, these have not been integrated into the tracking process. Our hypothesis is that a significant improvement can result from the use of context in dynamically influencing the linking of object detections, during the tracking process. To verify this hypothesis, we augment a state of the art dynamic programming based tracker with contextual information by reformulating the maximum a posteriori (MAP) estimation formulation. This formulation introduces contextual factors that first of all augment detection strengths and secondly provides temporal context. We allow both these types of factors to contribute organically to the linking process by learning the relative contribution of each of these factors jointly during a gradient decent based optimisation process. Our experiments demonstrate that the proposed approach contributes to a significantly superior performance on a recent challenging video dataset, which captures complex scenes with a wide range of object types and diverse backgrounds.**

I. Introduction

An important aspect of scene understanding is to detect and track objects, so that they may be used to model the behaviour of individual objects and their interactions in a scene. However, the task of obtaining correct detections often tends to come at the cost of a significant number of false positives, adding considerable noise to the subsequent tracking process.

This work is based on our premise that these false positives tend to be both spatially and temporally out of context. For example, spatially out-of-context false positives tend to be inconsistent with respect to the scene geometry in terms of its size or position, or may lack a distinct object-like-profile with respect to their immediate background. Temporally out-of-context false positives tend to be inconsistent with their immediate temporal neighbourhood.

Previous work has focussed on *solely* using spatial context during detection and temporal context during tracking respectively. Many detection based approaches have shown that the incorporation of spatial contextual information can significantly reduce the number of false positives. These approaches [1], [2] have concentrated on improving object detections by using different types of features such as gradient, colour and texture to represent objects, while also modelling the generic objectness cues [3]. Recent research [4], [5] has also demonstrated the advantage of using surface and view point as contexts. On the other hand, tracking in [6], [7] conceives the task as a separate procedure for linking the resulting detections and they incorporate temporal context to improve tracking performance.

Our work incorporates both spatial and temporal context into the tracking process. This way we simultaneously reduce false positives that are both spatially and temporally out-of-context. The proposed approach is evaluated on a publicly available challenging dataset consisting of 300 videos. These videos depict interactions between objects, most of them involving persons who perform common verbs in diverse and complex backgrounds. Our experimental results demonstrate that we can obtain tracks with a significant drop in the number of false positives, while only minimally reducing the number of true positives.

We organise the paper as follows. In Sec. II we briefly provide a literature survey on relevant research. We present the formulation of our approach in Sec. III. Our datasets and experiments are described in Sec. V. We conclude the paper with a summary and a description of future work in Sec. VI.

II. Related Work

Graphical models such as Markov random fields (MRF) and conditional random fields (CRF) have been used to include information about pixels surrounding a scanning-window detection and thus include the contextual information. The authors in [8] and [9] build a representation of context from low level features and use them to facilitate object detection. The authors in [10] propose *textons* that model shape and texture in a CRF to segment an image into semantic categories by exploiting context. Another approach is to combine both local and global contexts in a hierarchical field framework [11].

A related approach [12] obtains a 2D scene gist by computing global statistics of an image to capture the gist, providing the context of an object. Semantic scene context for object detection is exploited in [13] where object and scene categorisations are integrated. Spatial and co-occurrence relations between objects are used as contexts in [14] and [15].

Geometric scene structure such as surface and viewpoint have been shown to provide good object contexts. The authors in [5] integrate a multi-view object representation (using class specific deformable templates) as a likelihood with spatial features – surface, viewpoint and temporal features – foreground probability maps, local object trajectory predictions, as prior probabilities. Another approach is to model factors such as the interdependence of objects, surface and camera viewpoint and to use them in a iterative fashion in order to refine each other [4]. In recent work [16] combines various types of contexts for object detection within images.

Much previous research on tracking [7], [6] has posed the task as linking the observations. Contextual information has been used more for filtering the detections prior to tracking

or filtering the tracks post tracking, rather than influencing the linking during the tracking process itself. For example, the authors in [17] focus on "filtering" tracks using contextual information such as a viewpoint filter, foreground filter and trajectory-like filter. A few exceptions to this trend can be seen in [18], where the authors incorporate object-level spatio-temporal relationships as context using a MRF to improve tracking. Another such example is [19] which aims at improving tracking in videos using events as context.

However, there exists a significant gap between research on the role of context in object detection and their combination with temporal context within the tracking process for video analysis. We show that the simultaneous incorporation of appearance, geometric and temporal context can significantly improve the outcome of tracking.

III. Problem Formulation

We consider a video which is a time series of images $I = \{I^1, ..., I^t, ..., I^N\}$; we would like to interpret them with a non-overlapping set of tracks Γ. Each track $T \subseteq \Gamma$ is in turn composed of a temporally contiguous sequence of windows (detections), where the window X^t is one such at time t. For each window X^t we extract a set of local features L^t, the details of which are elaborated in the next section. For a given t, let $\mathcal{X}^t$ be the set of context windows e.g. sky and groundplane. The set of all features of all the windows in a track T are denoted by L and $\mathcal{X}$ is the set of all context windows for the track T, i.e $\{x : x \in \mathcal{X}^t, 1 \leq t \leq N\}$.

We formulate the task of context aware tracking as finding the optimal set of tracks $\hat{\Gamma}$ that maximizes the following product of joint probability distributions $P(T, L_T, \mathcal{X}, \Theta)$.

$$\hat{\Gamma} \quad = \quad \arg\max_{\Gamma} \prod_{T \in \Gamma} P(T, L_T, \mathcal{X}, \Theta)$$

We expand the joint probability distribution $P(T, L_T, \mathcal{X}, \Theta)$ as follows.

$$P(T, L_T, \mathcal{X}, \Theta) \approx \prod_{\substack{t \\ X^t \in T}} P(X^t, L^t | \mathcal{X}^t, X^{t-1:t-2}, L^{t-1}, \Theta)$$

$$\approx \prod_{\substack{t \\ X^t \in T}} P(X^t, L^t | \mathcal{X}^t, \Theta_{app}, \Theta_{spt}) P(X^t, L^t | \mathcal{X}^t, X^{t-1:t-2}, L^{t-1}, \Theta_{tem})$$

The first term $P(X^t, L^t | \mathcal{X}^t, \Theta_{app}, \Theta_{spt})$ leads to a *static understanding of the scene* for a particular frame t, where both appearance of objects and their spatial relations with their context is captured. We factorize this term as follows.

$$P(X^t, L^t | \mathcal{X}^t, \Theta_{app}, \Theta_{spt}) = P(L^t | X^t, \Theta_{app}) P(X^t | \mathcal{X}^t, \Theta_{spt})$$

The likelihood $P(L^t | X^t, \Theta_{app})$ models object appearance in terms of the probability of a feature vector L^t corresponding to a window X^t and an appearance model Θ_{app}. The second probability $P(X^t | \mathcal{X}, \Theta_{spt})$ models spatial context in terms of the probability of a window X^t given a set of windows $\mathcal{X}^t$ that form the spatial context of X^t using a model of spatial relationships Θ_{spt}.

The second term $P(X^t, L^t | \mathcal{X}^t, X^{t-1:t-2}, L^{t-1}, \Theta_{tem})$ leads to a *dynamic understanding of the scene* in terms of the relationship between an object and its temporal context i.e. between the objects corresponding to the previous two frames that are in the same track.

In the following paragraphs, we expand each of these terms that correspond to static and dynamic scene contexts respectively.

A. Object Appearance

We model the likelihood $P(L^t | X^t, \Theta_{app})$ for object appearance in terms of class specific features and scene appearance context as explained below.

$$P(L^t | X^t, \Theta_{app}) \quad = \quad \underbrace{P(L^t_{hog} | X^t, \Theta_{hog}) P(X^t | \Theta_{ar})}_{\text{class specific features}}$$
$$\underbrace{P(L^t_{cc} | X^t, \Theta_{cc}) P(L^t_{ed} | X^t, \Theta_{ed}) P(L^t_{ss} | X^t, \Theta_{ss})}_{\text{scene appearance context}}$$

1) ***Class Specific Features****:* We primarily use HOG features for representing the appearances of object classes and obtain trained part based models [20] for each object class. Another class specific feature is aspect ratio, which is useful since objects tend to have characteristic aspect ratios depending on the object class they belong to – people are more vertical and cars are more horizontal. Thus, a probability is computed with respect to HOG features $P(L^t_{hog} | X^t, \Theta_{hog}) = 1/z_2 \exp(\Theta_{hog} D(X^t))$ using the Boltzmann distribution[1] on the corresponding detection score $D(X^t)$. We model the probability with respect to aspect ratio $P(X^t | \Theta_{ar}) = \mathcal{N}\big(\chi(a(X^t), \Theta_{ar})\big)$ in terms of a Gaussian distribution.

2) ***Scene Appearance Context****:* One of the indicators of the presence of an object is its distinctiveness from its scene context. Three recently introduced measures [3] of an object's distinctiveness from its scene context are colour contrast, edge density and superpixel straddling. As we found the original measures in this work computationally expensive, we employ robust alternatives of these measures by defining an interior ring $\mathcal{R}_I(X^t)$ and an exterior ring $\mathcal{R}_E(X^t)$ with respect to the window X^t. The outer ring serves as an immediate context that can be compared to the inner ring. Using appropriate features in the inner and outer ring respectively, we express the likelihood for scene appearance context in terms of three probabilities corresponding to colour contrast $P(L^t_{cc} | X^t, \Theta_{cc})$, edge density $P(L^t_{ed} | X^t, \Theta_{ed})$ and superpixel straddling $P(L^t_{ss} | X^t, \Theta_{ss})$ as follows.

Objects tend to have a different appearance (colour distribution) to their immediate background contexts. The colour contrast probability $P(L^t_{cc} | X^t, \Theta_{cc})$ expresses the degree of discrepancy between the LAB histograms $h(\cdot)$ in the inner ring $\mathcal{R}_I(X^t)$ of a window X^t to another in the outer ring $\mathcal{R}_E(X^t)$.

[1]The probability of x is defined by assigning it to energy $E(x)$, which is converted into a probability using the Boltzmann distribution $P(x) = \frac{\exp(-E(x))}{\sum_x \exp(-E(x))}$

$$P(L^t_{cc}|X^t,\Theta_{cc}) = \frac{1}{z_3}\exp(-\Theta_{cc}\chi^{-2}(h(\mathcal{R}_I(X^t),h(\mathcal{R}_E(X^t)))$$

It is expected that objects tend to have well defined boundaries, characterised by the presence of areas of high edge densities near their enclosing bounding boxes. The edge density probability $P(L^t_{ed}|X^t,\Theta_{ed})$ is expressed in terms of the degree of edge density in the inner ring $\mathcal{R}_I(X^t)$. Accordingly, we count the number of pixels p for which the binary edgemap $I_{ed}(p)$ obtained using the Canny detector gives a value of one within the inner ring $\mathcal{R}_I(X^t)$. Then we normalize this count by the perimeter $\|\mathcal{R}_I(X^t)\|$ of this ring.

$$P(L^t_{ed}|X^t,\Theta_{ed}) = \frac{1}{z_4}\exp\Big(-\frac{\Theta_{ed}\|\mathcal{R}_I(X^t)\|}{\sum_{p\in\mathcal{R}_I(X^t)} I_{ed}(p)}\Big)$$

Superpixels straddling measures the extent to which the superpixels of the image within a detection bounding box straddle the bounding box, as the superpixels have the property that even though they over segment an object, they preserve object boundaries. That is, most pixels in a superpixel belong to the same object and hence do not tend to straddle these boundaries, or the corresponding bounding boxes. For each superpixel s obtained with the segmentation scale Θ_{ss}, we express superpixel straddling probability $P(L^t_{ss}|X^t,\Theta_{ss})$ in terms of the sums of the ratio of the area $R_I(X^t,s)\cup X^t$ in the interior ring $R_I(X^t,s)$ to the area in the exterior ring $R_E(X^t,s)\cup X^t$, across the set of superpixels $s\in SP(X^t)$ for the detection X^t.

$$P(L^t_{ss}|X^t,\Theta_{ss})=\frac{1}{z_5}\exp\Big(-\Theta_{ss}\sum_{s\in SP(X^t)}\frac{\mathcal{R}_I(X^t,s)\cup X^t}{\mathcal{R}_E(X^t,s)\cup X^t}\Big)$$

B. *Spatial Context*

The locations of objects are naturally constrained to specific surfaces, and their sizes are constrained by the distance from the camera. Therefore, we model the likelihood $P(X^t|\mathcal{X}^t,\Theta_{spt})$ for spatial context in terms of a scene geometric context [2]. Hence, we use surface and viewpoint as scene geometric context and obtain the corresponding likelihoods $P(X^t|X^{'},\Theta_{sf})$ and $P(X^t|X^{''},\Theta_{vp})$ as follows.

$$P(X^t|\mathcal{X}^t,\Theta_{spt}) = P(X^t|X^{'},\Theta_{sf})P(X^t|X^{''},\Theta_{vp})$$

1) ***Surface:*** To obtain the approximate 3D surface orientations in the image, we apply the method of [22], which produces confidence maps for three main classes: ground, vertical, and sky. The objects under consideration in our dataset primarily lie on the ground. Hence $P(X^t|X^{'},\Theta_{sf})$ represents the spatial relationship between X^t and the window of the ground plane $X^{'}$ and is computed in terms of the average of ground surface probabilities $G(p)$ across all pixels $p\in S(X^t,X^{'})$ where S provides the set of all shared pixels between X^t and $X^{'}$.

$$P(X^t|X^{'},\Theta_{sf}) = \frac{1}{z_2}\exp\Big(\frac{-\Theta_{sf}\|S(X^t,X^{'})\|}{\sum_{p\in S(X^t,X^{'})}G(p)}\Big)$$

2) ***Viewpoint:*** In order to model the view point likelihood $P(X^t|X^{''},\Theta_{vp})$, we obtain the depth $d(X^t)$ of a window X^t using a pre-defined homography mapping, which we assume to be constant for all the videos[3]. We estimate a window $X^{''}$ based on the expected size of X^t at the depth $d(X^t)$ in the image plane. The view point likelihood is then modelled in terms of a Gaussian distribution over a ratio of the estimated height $h^{'}(X^{''})$ and the observed height $h(X^t)$.

$$P(X^t|X^{''},\Theta_{vp}) = \mathcal{N}\Big(\frac{h^{'}(X^{''})}{h(X^t)},\Theta_{vp}\Big)$$

C. *Temporal Context*

The temporal context is simply the temporal neighbourhood of each detection with respect to the track. We factorize the temporal context term $P(X^t,L^t|\mathcal{X}^t,X^{t-1:t-2},L^{t-1},\Theta_{tem})$ as follows.

$$P(X^t,L^t|\mathcal{X}^t,X^{t-1:t-2},L^{t-1},\Theta_{tem}) = P(X^t|X^{t-1:t-2},\Theta_{mcont})P(L^t|L^{t-1},\Theta_{acont})$$

The first probability $P(X^t|X^{t-1:t-2},\Theta_{mcont})$ models temporal context in terms of motion continuity i.e. the probability of a window X^t given two previous windows $X^{t-1:t-2}$. We model motion continuity in terms of the similarity in magnitude and directions between the displacement vectors $v_1=\overrightarrow{b_{n-2}b_{n-1}}$ and $v_2=\overrightarrow{b_{n-1}b_n}$ corresponding to the bounding boxes b of the windows in three consecutive frames of a track, X^{t-2},X^{t-1} and X^t respectively. We model the motion continuity likelihood as:

$$P(X^t|X^{t-1:t-2},\Theta_{mcont})=\frac{1}{z_7}\exp\Big(-\Theta_p\big(\frac{v_1\cdot v_2}{\|v_1\|\|v_2\|}+\frac{\sqrt{v_1\cdot v_2}}{\|v_1\|+\|v_2\|}\big)^{-1}\Big)$$

The second probability $P(L^t|L^{t-1},\Theta_{acont})$ models temporal context in terms of appearance continuity i.e. the probability of a window's appearance features L^t given that of a previous window i.e. L^{t-1}. We model appearance continuity in terms of chi-squared similarity between the LAB histograms h^t_{lab} and h^{t-1}_{lab} of two windows X^t and X^{t-1} in consecutive frames of a track, respectively. Therefore we compute colour continuity as:

$$P(L^t|L^{t-1},\Theta_{acont}) = \frac{1}{z_6}\exp\big(-\Theta_c\chi^{-2}(h^t_{lab},h^{t-1}_{lab})\big).$$

IV. Gradient Descent Optimisation

Let $\hat{\Gamma}(\Theta)$ be an optimal track hypothesis returned by the tracker, with respect to a weight vector Θ, where Θ defines the contribution of each type of contextual features to the optimal tracks. Given a set of tracks Γ_{GT} in the ground truth, the objective of parameter learning is to search for the optimal $\hat{\Theta}$ that gives minimum error between the hypothesised track $\hat{\Gamma}(\Theta)$ and the ground truth track Γ_{GT}. In other words,

$$\hat{\Theta} = \arg\min_{\Theta}\mathcal{D}(\hat{\Gamma}(\Theta),\Gamma_{GT})$$

[2] In [21] we incorporate another spatial context, specifically event context, which is obtained in terms of the spatial relationship between people and objects for a given event type *carry*.

[3] We have observed that the camera viewpoint parameters are similar for most videos in the dataset used in our experiments. In the future, we plan to estimate the homography for any new video during test time.

where $\mathcal{D}$ is a function that measures the divergence between a set of tracks and the ground truth for a video, defined between two sets of tracks Γ and Γ' as

$$\mathcal{D}(\Gamma,\Gamma') = \sum_{\substack{T\in\Gamma:\nexists T'\in\Gamma' \\ :T=H(T')}} \|Vol(T)\| - \sum_{\substack{T'\in\Gamma':\exists T\in\Gamma \\ :T=H(T')}} \cap\,(Vol(T), Vol(H(T)))$$

where $H:\Gamma'\rightarrow\Gamma$ is the optimal assignment between Γ' to Γ, computed using bipartite matching [23] based on the overlap of their respective volumes. The volume of a track is computed from the three dimensional tube traced by its bounding box through time. The first term measures the sum of volumes of tracks in Γ, none of which are assigned to any track in Γ' by the assignment H. The second term measures the sum of volume intersections between pairs of tracks $T\in\Gamma$ and $T'\in\Gamma'$, such that the track T' is assigned to track T by H.

We apply a basic gradient descent based optimisation process with line search [24] to find the best solution $\hat{\Theta}$. Accordingly, we define $J(\Theta)=\mathcal{D}(\Gamma(\Theta),\Gamma_{GT})$ as the criterion function, and $\Theta(1)$ the initial weight vector with some arbitrary values. At each step, we compute the gradient vector $\nabla J(\Theta(k))$, and the next value $\Theta(k+1)$ is obtained by moving some distance toward the steepest direction, i.e. along the negative of gradient. The update process can be defined as

$$\Theta(k+1)=\Theta(k)-\eta(k)\nabla J(\Theta(k))$$

where $\eta(\cdot)$ is the learning rate, a positive scalar factor that reflects to the degree of change at each step. The processes repeats until the termination condition is satisfied, that is, $|\eta(k)\nabla J(\Theta(k))|<\lambda$, where λ is a predefined threshold for the purpose of convergence.

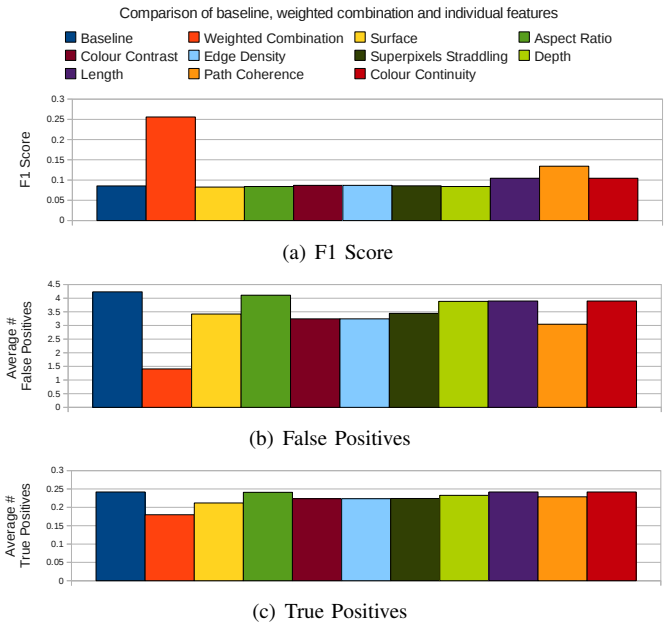
V. Experiments

In the following, we first describe the dataset and experimental setups. We then present a quantitative evaluation followed by a qualitative explanation using image sequences that convey the success and limitations of the proposed approach.

A. Dataset and Experimental Setup

We evaluated the proposed approach using the year-one (Y1) corpus produced by DARPA for the Minds Eye program [25], which consists of 300 videos. These videos are provided at 30 frames per second and range between 40 to 1500 frames, with an average of 438 frames and a resolution of 1280×720 pixels. They depict interactions between a variety of objects, most of them involving people who perform common verbs such as approaching and exchanging. Human-annotated tracks for people and cars have been publicly made available by Stanford University. We used these human-annotated tracks as ground truth to evaluate and compare the tracking performances obtained using the proposed approach with that of the baseline tracker. Before we evaluate our tracker, we randomly sample 10 videos from which we learn the parameters $\Theta_{ar},\Theta_{cc},\Theta_{ed},\Theta_{ss},\Theta_{vp}$. We then remove these videos from the dataset and henceforth, refer to the rest of the videos as our dataset for the purpose of evaluation.

We first ran an "off the shelf" object detector [20] for the car and person class with a low non-maximum suppression threshold, generating around 50 detections per frame per class



(a) F1 Score

(b) False Positives

(c) True Positives

Fig. 1. The results of the baseline, weighted combination and for each of the contextual features are shown. The chart (a) shows the F1 score, (b) the average false positives and (c) average true positives across all videos.

for each video in the entire dataset. Then, we adopted a five-fold video based cross validation scheme for evaluation where we partitioned the videos into 5 random folds. We trained the proposed *context aware tracker* using the scheme described in Sec. IV where we trained on each of the folds and applied this tracker on the rest of the corresponding four folds for testing. To the mentioned detections, we also applied an "off the shelf" state of the art tracker [26] to serve as a baseline against which we compare our performance using our approach.

We use a standard F1 score as a scoring criterion to evaluate and compare the performance of the proposed tracker with that of the baseline tracker. More specifically, the F1 score for tracking is obtained by first computing an optimal assignment between human-annotated tracks and automatic tracks based on the extent that they overlap. The overlap between the tracks is computed as the volume of their intersection divided by the union of the volumes between the two sets of tracks. The optimal assignment is computed using bipartite matching [23] and we consider only those assignments whose overlap is more than 50%. The optimal assignment between the human-annotated tracks and the automatic tracks is used to compute the F1 scores.

B. Quantitative Evaluation

To assess the role of context in tracking, we present the performance of the context aware tracker along with the baseline tracker which does not use context. To understand the role of each of the context features individually and to compare their performances with the weighted combination of the context aware tracker, we also present the performances of the context aware tracker using just one feature at a time. We describe the overall performance in terms of the average F1 scores across five folds in Fig. 1(a). A more detailed perspective is gained by examining the TP and FP scores

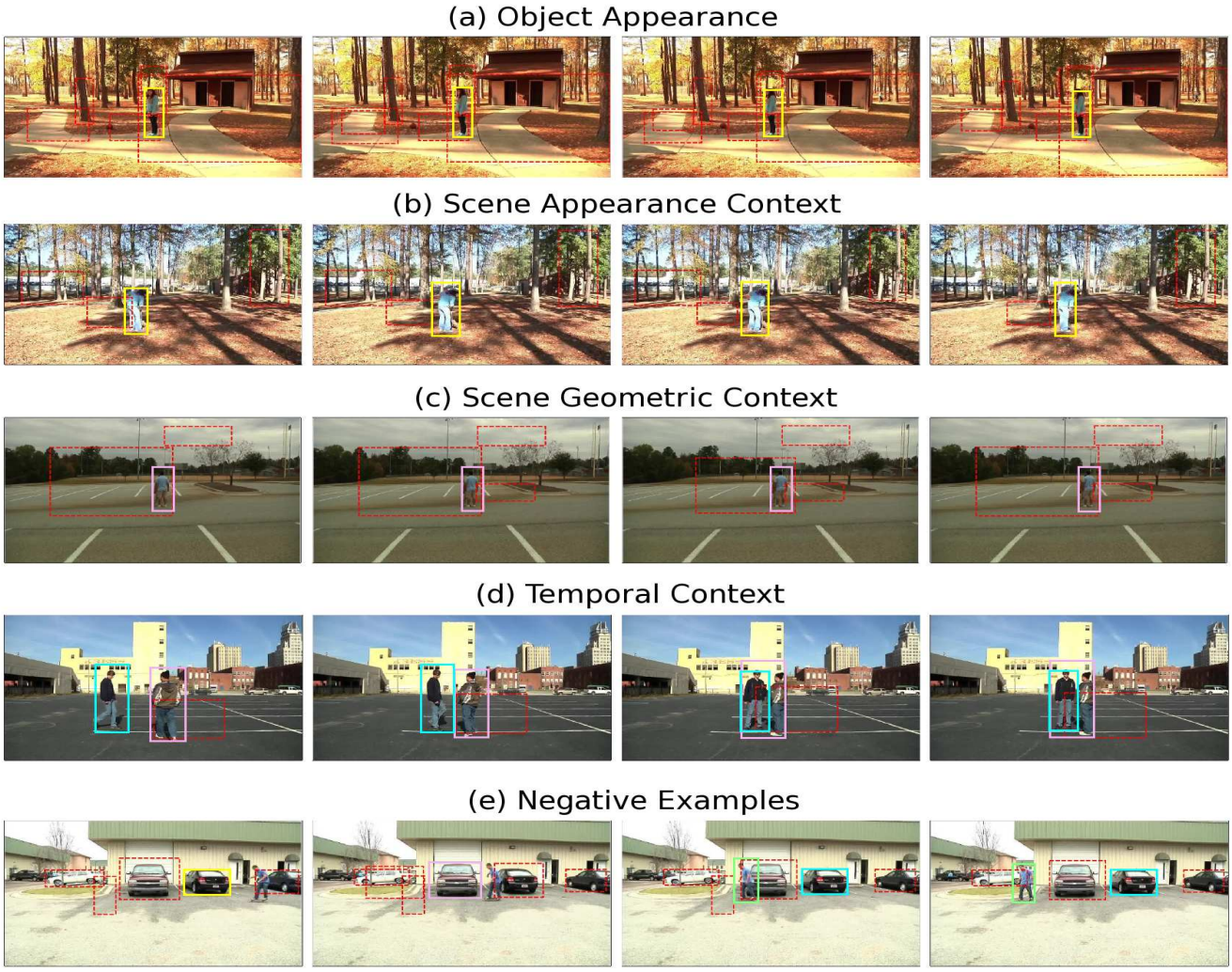


Fig. 2. This figure illustrates the qualitative performance of the proposed framework by presenting a set of continuous images for positive cases in each feature type in (a)–(d) while also providing negative examples in (e). Each row of images consists of a set of tracks which are represented by two types of bounding boxes, namely solid and dotted. The set of all bounding boxes represent the original tracks while the solid and dotted tracks represent the tracks accepted and rejected by our framework respectively.

averaged across the test videos in five folds and across two classes for which ground truth is available - persons and cars.

From Fig. 1(a) it can be concluded that the proposed approach of incorporating contexts into tracking improves the overall F1 scores by 0.17, with respect to the baseline tracker. Moreover, Fig. 1(b) and 1(c) show that while the proposed approach significantly reduces the average number of false positives by 2.82, the average number of true positives decreases minimally. We can also conclude that the learned-weighted combination of contextual features achieves a significantly superior performance in terms of F1 scores and average FP over each of these features only when they are used individually. We have observed that if each of the contextual features were to be used separately in a preprocessing step to filter out detections, altogether, they tend to remove most of the true positives. However, their weighted combination is successful in retaining most of the true positives, removing only a minimal amount (0.07). The next section provides a qualitative explanation for the minimum decrease in true positives.

C. Qualitative Evaluation

We now turn to a qualitative explanation of the above empirical results, where we first present sequences of image samples from different videos that illustrate the role of each of the contextual factors where they have been individually successful in reducing false positives. It is interesting to observe that in Fig. 2 (a), all car tracks have been rejected due to low strengths in detection using class specific features, while one of the person tracks has been retained as a true positive due to the relatively high detection strength. In Fig. 2(b) the tracks whose bounding boxes do not capture any relevant objects, e.g. amorphous surfaces such as the ground and porous surfaces like the trees, have been rejected based on the generic objectness features. For the same video sequence, the track that captures the person has been correctly retained. It can be observed from Fig. 2(c), tracks consisting of detections whose positions and sizes are inconsistent with being on a certain surface and depth, e.g. car detections in the sky, or very large car detections at a considerable depth from the camera have been successfully rejected. Again, we note that

the person track has been correctly retained. By considering the movement of the bounding box in Fig. 2 (d), it can be observed that it is rejected on the basis of a lack of continuity in motion and appearance.

While the overall performance of the context aware tracker is significantly superior to the baseline and to using each feature separately, there is however a small 0.07 decrease in the average number of true positives. An analysis of test videos where there is a decrease in true positives suggests that there are certain *characteristic videos*, and where this decrease tends to occur due to certain *typical reasons*. The two main reasons that we have identified are as follows. First, we have observed that due to the object interactions that are present in these videos, sometimes the appearance and motion features tend to change substantially during inter-object occlusions. Under such circumstances, the coherence measures of motion and appearance tend to give low scores. Fig. 2(e) illustrates such a case, where a person occludes a car affecting its appearance profile, resulting in the loss of a potential true positive. The second reason is where one of the objects acts as a vertical surface for another, causing it to appear as though the base of the bounding box of the latter object is on a vertical surface. This can have an effect of reducing the surface scores, thus removing a potential true positive.

VI. Summary and Future Work

In this work, we have considered four types of context - object appearance context, scene geometric context, scene appearance context and temporal context. We have incorporated these four factors into the tracking process by means of a MAP formulation. This formulation introduces contextual factors that first of all augment detection strengths and secondly provide temporal context. We allow both these types of factors to contribute organically to the linking process by optimizing the relative contribution of each of these factors jointly during the parameter learning phase. The results presented in this paper use a state of the art tracker and show that by using a weighted combination of contextual information the performance of the original tracker is improved. We believe that the same approach can be applied to other trackers in order to improve their performance, although this requires further analysis. A qualitative examination of results clearly demonstrates the role of each of the contextual factors in improving the performance of tracking.

Several issues remain to be explored in the future. In this work, we have used three simple surface classes as spatial scene contexts. In the future we plan to include a more fine grained interpretation of the scene in terms of categories such as road and grass. Also, we plan to model the spatial relations between the objects as context. We believe that this would further improve the performance of our approach by addressing the problems discussed above. We are currently exploring the incorporation of event context into the tracking process [21].

Acknowledgment

The financial support of DARPA Mind's Eye project VIGIL (W911NF-10-C-0083) and the EU projects RACE (FP7-ICT-287752) and STRANDS (FP7-ICT-600623) is gratefully acknowledged.

References

[1] B. Wu and R. Nevatia, "Optimizing discrimination-efficiency tradeoff in integrating heterogeneous local features for object detection." in *CVPR*, 2008.

[2] F. Han, Y. Shan, H. Sawhney, and R. Kumar, "Discovering class specific composite features through discriminative sampling with Swendsen-Wang Cut," *CVPR*, 2008.

[3] B. Alexe, T. Deselaers, and V. Ferrari, "What is an object?" *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2010.

[4] D. Hoiem, A. A. Efros, and M. Hebert, "Putting objects in perspective," *International Journal on Computer Vision*, vol. 80, no. 1, pp. 3–15, Oct. 2008.

[5] X. Liu, L. Lin, S. Yan, H. Jin, and W. Tao, "Integrating spatio-temporal context with multiview representation for object recognition in visual surveillance," *Circuits and Systems for Video Technology,*, vol. 21, no. 4, pp. 393 –407, april 2011.

[6] H. Yang, L. Shao, F. Zheng, L. Wang, and Z. Song, "Recent advances and trends in visual tracking: A review." *Neurocomputing*, vol. 74, no. 18, pp. 3823–3831, 2011.

[7] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Comput. Surv.*, vol. 38, no. 4, Dec. 2006.

[8] A. Torralba, "Contextual Priming for Object Detection," *International Journal on Computer Vision*, vol. 53, no. 2, pp. 169–191, Jul. 2003.

[9] L. Wolf and S. Bileschi, "A critical view of context," *International Journal on Computer Vision*, vol. 69, no. 2, pp. 251–261, Aug. 2006.

[10] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost: Joint appearance, shape and context modeling for multi-class object," in *ECCV*, 2006, pp. 1–15.

[11] S. Kumar and M. Hebert, "A hierarchical field framework for unified context-based classification," in *ICCV*, 2005, pp. 1284–1291.

[12] A. Torralba, A. Oliva, and W. T. Freeman, "Object recognition by scene alignment," *Journal of Vision*, 2007.

[13] L.-J. Li and L. Fei-Fei, "What, where and who? classifying events by scene and object recognition," in *IEEE*, oct. 2007, pp. 1 –8.

[14] C. Galleguillos, A. Rabinovich, and S. Belongie, "Object categorization using co-occurrence, location and appearance," in *CVPR*, june 2008, pp. 1–8.

[15] G. Heitz and D. Koller, "Learning Spatial Context: Using Stuff to Find Things," in *ECCV*. Springer, 2008, vol. 5302, ch. 4, pp. 30–43.

[16] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Hebert, "An empirical study of context in object detection," in *CVPR*, June 2009.

[17] S. Stalder, H. Grabner, and L. Van Gool, "Cascaded confidence filtering for improved tracking-by-detection," in *ECCV*. Springer, 2010, pp. 369–382.

[18] Y. Li and R. Nevatia, "Key object driven multi-category object recognition, localization and tracking using spatio-temporal context," in *ECCV*. Springer, 2008, pp. 409–422.

[19] A. Barbu, A. Michaux, S. Narayanaswamy, and J. M. Siskind, "Simultaneous object detection, tracking, and event recognition," *CoRR*, vol. abs/1204.2741, 2012.

[20] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *PAMI*, vol. 32, no. 9, pp. 1627–1645, 2010.

[21] A. Tavanai, M. Sridhar, F. Gu, A. Cohn, and D. Hogg, "Carried object detection and tracking using geometric shape models and spatio-temporal consistency," in *Computer Vision Systems*, ser. LNCS. Springer, 2013, vol. 7963, pp. 223–233.

[22] D. Hoiem, A. A. Efros, and M. Hebert, "Geometric context from a single image," in *In ICCV*, 2005, pp. 654–661.

[23] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics (NRL)*, vol. 52, no. 1, pp. 7–21, 2005.

[24] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. Wiley-Interscience, Nov. 2001.

[25] DARPA. (2011) Mind's eye challenge http://www.visint.org/. [Online]. Available: http://www.visint.org/

[26] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes, "Globally-optimal greedy algorithms for tracking a variable number of objects," in *CVPR*, 2011, pp. 1201–1208.