FaceHugger: The ALIEN Tracker Applied to Faces

Federico Pernici

MICC - University Of Florence, Italy

Abstract. This paper proposes an online tracking method which has been inspired by studying the effects of Scale Invariant Feature Transform (SIFT) when applied to objects assumed to be flat even though they are not. The consequent deviation from flatness induces *nuisance factors* that act on the feature representation in a manner for which no general local invariants can be computed, such as in the case of occlusion, sensor quantization and casting shadows. However, if features are over-represented, they can provide the necessary information to build online, a robust object/context discriminative classifier. This is achieved based on weakly aligned *multiple instance* local features in a sense that will be made clear in the rest of this paper. According to this observation, we present a non parametric online tracking by detection approach that yields state of the art performance.

Specific tests on video sequences of faces show excellent long-term tracking performance in unconstrained videos.

1 Introduction

Tracking is a fundamental problem in computer vision. Several aspects of this difficult task have been considered in literature. Generally speaking, difficulties arise depending on the type of information that have to be tracked: 3D pose, imaged 2D location, imaged 2D shape, 3D shape, imaged 2D articulated body shape, 3D articulated body shape, etc. (see [1] for a review and a classification). Besides dealing with the inherent difficulties related to the specific information of interest, effective methods must also provide robust object representation coping with nuisance factors that affect the image formation process. Illumination, viewpoint, shadows, occlusion and clutter have indeed little to do[1] with the tracking of the physical quantities we are interested in. Further complexity is generated by objects or cameras themselves. For example objects may have non-rigid shape such as in the case of faces or may be made of translucent or reflective materials and camera sensors may suffer from the effects of noise, sensor quantization and motion blur.

In addition to these intrinsic problems, practical requirements such as: 1) long-term tracking; 2) object reacquisition after total occlusion and 3) the amount of partial occlusion at which to successfully track an object, may hinder the accomplishment of the tracking task. In some applications, the object to be tracked is known in advance and it is possible to incorporate specific prior knowledge when designing the tracker to alleviate some of these issues [2]. However, the general case of tracking arbitrary objects by simply specifying a *single (one-shot) training example at runtime*, is a challenging

[1] Indeed their relationships are too complex to be estimated.

A. Fusiello et al. (Eds.): ECCV 2012 Ws/Demos, Part III, LNCS 7585, pp. 597–601, 2012.

open problem which deserves particular attention. In this scenario, the tracker must be able to model the appearance of the object on-the-fly by generating and labeling image features and learning the model of the object appearance. This basic formulation naturally leads to the semi-supervised learning procedure.

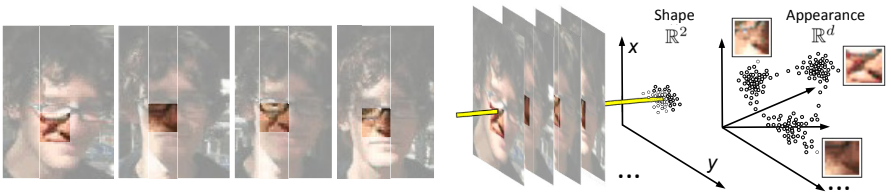
2 Related Work

Despite all the difficulties we introduced so far, a number of methods has been developed in which tracking is considered as simple as 2D image bounding box localization and what is really tracked is indeed the non-stationary image appearance of the object, irrespective of its imaged 3D physical quantities: [3–8].

In a recent quantitative comparison [10], among others, three methods emerged distinguishing for their positive performance and for their algorithmic design and image representation peculiarity: [5, 7, 8]. Their main differences rely on how they consider *the template update problem* which primarily impacts on the drift of the tracker [11]. Babenko et al. [8] address the problem by building an evolving boosting classifier that tracks bags of image patches. Kalal et al.[7] combine a optic flow tracker with an online random forest as introduced in [6]. In Mei and Ling [5] the tracking problem is formulated as finding a sparse representation of the object candidate combining trivial templates which are primarily responsible for the presence and the absence of certain object image patches. We argue that positive performance is intrinsically in the multiview appearance representation which allows overcoming the feature invariance and/or feature selection based on machine learning methods. MILTrack [8], for example, adopting bag of image patches can cope for misalignment and occlusion by adding novel examples as new instances for object representation. Based on this general observation, we propose a technique principally motivated by local feature invariance and by the underlying image formation process. It comprises multiple instances of local features combined with a global shape prior, expressed in terms of a 2D similarity transformation and it approximates object surfaces as nearly planar for which SIFT matching (or other local scale invariant features) has proven to be effective in the solution of the problem [12]. Conscious of the limits of local features invariance, 3D shape deviations from planarity and their interactions with shadow and occlusion are (over)-represented through multiple instances of the same features after a weak alignment along the object template (see Fig.1(a) and 1(b)). For this motivation we call our method ALIEN, Appearance Learning In Evidential Nuisance, since it is based on the physical observation that if the object is reasonably convex, known critical nuisance factors which cannot be neutralized, can be managed based on multiple instances of features selected and updated according to a weak global shape model. This novel representation is exploited in a discriminative background/foreground online tracking (by detection) method which performs feature selection and feature update. The resulting technique allows tracking to continue under severe visibility artifacts. In our demo we build on the ALIEN method to develop a face tracking application in which face re-detection is exploited to distinguish face identities when objects move in and out of the field of view. We call our application FaceHugger, since it "sticks" firmly to the face even in unrestricted viewing conditions.

3 The ALIEN Tracker

Given a bounding box defining an object of interest, our goal is to automatically and unambiguously determine which image features are the most useful in discriminating between the object and the rest of the imaged scene. The main components of our method are two nearest neighbor classifiers (NN); one for the object under tracking and the other for its context. The two classifiers are non-parametric defined in terms of the set of visual features they represent. The object classifier $\mathcal{T}_t$ represents object shape and appearance at time t by a number of features $N_{\mathcal{T}}$ as: $\mathcal{T}_t = \{(\mathbf{p}_i, \mathbf{d}_i)\}_{i=1}^{N_{\mathcal{T}}}$, where $\mathbf{p} \in \mathbb{R}^2$ is a point location in the object reference template with its associated image patch descriptor $\mathbf{d} \in \mathbb{R}^n$ as illustrated in Fig. 1(b). The second classifier $\mathcal{C}_t$ defines



(a) (b)

Fig. 1. The weakly aligned multi-instance local features concept in the case of tracking a face. (a): Four frames from the *trellis-sequence* [3] with highlighted appearance variations in a particular object region susceptible to self-occlusions and shadows. (b): Region representation after weak alignment. Feature locations describing 2D shape in the xy-coordinate system of the object template are shown with their associated appearance descriptors (128D).

the contextual appearance surrounding, in space and time, the object and is composed of only the appearance component (i.e. standard bag of features representation): $\mathcal{C}_t = \{\mathbf{d}_i\}_{i=1}^{N_{\mathcal{C}}}$, where $N_{\mathcal{C}}$ is the number of features and $\mathbf{d} \in \mathbb{R}^n$ is the associated visual descriptor. We use SIFT [12] as the features for both the classifiers, however any scale invariant representation can be plugged in. The final object detector, that will be detailed elsewhere, is composed by the tight interplay between the sets $\mathcal{T}_t$, $\mathcal{C}_t$ and the object state $\mathbf{x}_t$. The detector returns $p(y = 1|\mathcal{S}_t)$ where $\mathcal{S}_t = \{(\mathbf{p}_i, \mathbf{d}_i)\}_{i=1}^{N_{\mathcal{S}}}$ is the set of features extracted from an image search area $\mathbf{S}_t$ and y is a binary variable indicating the presence or the absence of the object of interest in that image region. Detector response is evaluated with a greedy strategy, to also obtain the tracker state. The tracker state $\mathbf{x}_t$ at time t includes the parameters to specify imaged object center location (x_t, y_t), scale s_t and the rotation angle θ_t with respect to the initial bounding box provided at time $t = 0$. Once the tracker state is estimated, we proceed to update the object/context appearance model. To this aim local features inside the Oriented Bounding Box OBB$(\hat{\mathbf{x}}_t)$ region, defined by the tracker state, are labeled as belonging to the object. While for context, we use the features belonging to the annular region surrounding the object accumulated over a time window of length l. Suppose that the classifier is evaluating its response in the estimated search area $\mathbf{S}_t$ at time t, our goal is to perform object detection and object appearance update using the representation we introduced. To this end, the following

three points are explicitly addressed by the method and detailed elsewhere for lack of space: (1) *Feature distinctiveness*. Descriptors alone are ambiguous because they can be interpreted as a valid description for both the object and its surround context. An analogous effect is produced by the inherent shape limit of the bounding box. (2) *Not up to date appearance*. Appearance must be updated according to the novel information provided by the detected object in the current image. (3) *Occlusion*. Occlusion must be detected in order to avoid updating the wrong appearance contaminating the object template.

4 Experimental Results: ALIEN vs. Predator

ALIEN was compared with results reported in the recent developed PREDATOR[2], which reports on performance of 5 trackers: Online Boosting (OB) [13], Semisupervised Boosting (SB) [14], Beyond Semisupervised (BS) [15], MIL [8] and CoGD [4] on 9 sequences. The sequences include full occlusion and two of them contain about 10000 frames. Performance are dominated by ALIEN and PREDATOR [7] which are designed for object reacquisition. As in [7], the performance was assessed using the Pascal Score and Table 1 shows the Precision, Recall, F-measure results. ALIEN achieved the best score in the sequences and matched the performance of the current state of the art method [7].

Table 1. ALIEN in comparison to results reported in [7] (Precision/Recall/F-measure). Bold numbers indicate the best score, italic numbers indicate the second best.

Sequence	Frames	OB [13]	SB [14]	BS [15]	MIL [8]	CoGD[4]	TLD [7]	ALIEN
David	761	0.01 / 0.01 / 0.01	0.27 / 0.27 / 0.27	0.16 / 0.12 / 0.13	0.06 / 0.06 / 0.06	0.99 / 0.99 / 0.99	**1.00 / 1.00 / 1.00**	*0.99 / 0.98 / 0.99*
Jumping	313	0.41 / 0.04 / 0.08	0.14 / 0.08 / 0.10	0.06 / 0.05 / 0.05	0.37 / 0.37 / 0.37	**1.00 / 0.99 / 1.00**	*0.99 / 0.99 / 0.99*	0.99 / 0.87 / 0.92
Pedestrian 1	140	0.36 / 0.09 / 0.14	0.20 / 0.14 / 0.16	0.10 / 0.04 / 0.05	0.42 / 0.42 / 0.42	0.99 / 0.99 / 0.99	**1.00 / 1.00 / 1.00**	**1.00 / 1.00 / 1.00**
Pedestrian 2	338	0.74 / 0.12 / 0.21	0.55 / 0.46 / 0.50	1.00 / 0.02 / 0.04	0.10 / 0.12 / 0.11	0.71 / 0.90 / 0.79	*0.89 / 0.92 / 0.91*	**0.93 / 0.92 / 0.93**
Pedestrian 3	184	1.00 / 0.33 / 0.49	0.41 / 0.33 / 0.36	0.81 / 0.40 / 0.54	0.49 / 0.58 / 0.53	0.84 / 0.99 / 0.91	**0.99 / 1.00 / 0.99**	*1.00 / 0.90 / 0.95*
Car	945	0.89 / 0.57 / 0.69	1.00 / 0.67 / 0.80	0.99 / 0.56 / 0.72	0.11 / 0.12 / 0.11	0.91 / 0.92 / 0.91	*0.92 / 0.97 / 0.94*	**0.95 / 1.00 / 0.98**
Motocross	2665	0.13 / 0.00 / 0.00	0.01 / 0.00 / 0.00	0.14 / 0.00 / 0.00	0.02 / 0.01 / 0.01	0.80 / 0.26 / 0.39	**0.67 / 0.58 / 0.62**	*0.49 / 0.58 / 0.54*
Volkswagen	8576	0.04 / 0.00 / 0.00	0.00 / 0.00 / 0.00	0.00 / 0.00 / 0.00	0.26 / 0.03 / 0.05	0.41 / 0.03 / 0.06	*0.54 / 0.64 / 0.59*	**0.99 / 0.70 / 0.82**
Carchase	9928	0.73 / 0.03 / 0.05	0.79 / 0.04 / 0.08	0.38 / 0.09 / 0.14	0.49 / 0.03 / 0.05	0.87 / 0.04 / 0.08	*0.50 / 0.40 / 0.45*	**0.73 / 0.68 / 0.70**
mean	-	0.40 / 0.10 / 0.15	0.30 / 0.18 / 0.20	0.33 / 0.11 / 0.15	0.21 / 0.15 / 0.15	0.68 / 0.55 / 0.55	*0.68 / 0.68 / 0.68*	**0.73 / 0.69 / 0.71**

5 Conclusion

In this paper, we have presented the main features of a method to track an unknown object in long video sequences under complex interactions between illumination, occlusion and object/camera motion. A real-time implementation of the framework has been evaluated under a publicly available dataset with an extensive set of experiments. Superiority of our approach with respect to state of the art methods was reported.

[2] Tracking-Learning-Detection (TLD) tracker [7] has been advertised under the name Predator.

References

1. Yilmaz, A., Javed, O., Shah, M.: Object tracking: A survey. ACM Computing Surveys 38, 13 (2006)
2. Lepetit, V., Fua, P.: Monocular model-based 3d tracking of rigid objects. Found. Trends. Comput. Graph. Vis. 1, 1–89 (2005)
3. Ross, D.A., Lim, J., Lin, R.S., Yang, M.H.: Incremental learning for robust visual tracking. Int. J. Comput. Vision 77, 125–141 (2008)
4. Yu, Q., Dinh, T.B., Medioni, G.G.: Online Tracking and Reacquisition Using Co-trained Generative and Discriminative Trackers. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 678–691. Springer, Heidelberg (2008)
5. Mei, X., Ling, H.: Robust visual tracking using l1 minimization. In: ICCV 2009, pp. 1436–1443 (2009)
6. Santner, J., Leistner, C., Saffari, A., Pock, T., Bischof, H.: Prost: Parallel robust online simple tracking. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 723–730 (2010)
7. Kalal, Z., Matas, J., Mikolajczyk, K.: P-n learning: Bootstrapping binary classifiers by structural constraints. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2010)
8. Babenko, B., Yang, M.H., Belongie, S.: Robust object tracking with online multiple instance learning. IEEE Transactions on Pattern Analysis and Machine Intelligence 33, 1619–1632 (2011)
9. Dinh, T.B., Vo, N., Medioni, G.: Context tracker: Exploring supporters and distracters in unconstrained environments. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2011)
10. Wang, Q., Chen, F., Xu, W., Yang, M.H.: An experimental comparison of online object tracking algorithms. In: Proceedings of SPIE: Image and Signal Processing Track (2011)
11. Matthews, I., Ishikawa, T., Baker, S.: The template update problem. In: Proceedings of the British Machine Vision Conference (2003)
12. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60, 91–110 (2004)
13. Grabner, H., Bischof, H.: On-line boosting and vision. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 260–267 (2006)
14. Grabner, H., Leistner, C., Bischof, H.: Semi-supervised On-Line Boosting for Robust Tracking. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 234–247. Springer, Heidelberg (2008)
15. Stalder, S., Grabner, H., Van Gool, L.: Beyond semi-supervised tracking: Tracking should be as simple as detection, but not simpler than recognition. In: OLCV 2009: 3rd On-line learning for Computer Vision Workshop (2009)