TEC2011-25995 EventVideo (2012-2014)

Strategies for Object Segmentation, Detection and Tracking in Complex Environments for Event Detection in Video Surveillance and Monitoring

D4.1

TRACKING (IN DENSE/CLUTTER ENVIRONMENTS)

Video Processing and Understanding Lab

Escuela Politécnica Superior

Universidad Autónoma de Madrid

Supported by GOBIERNO DE ESPAÑA MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD

AUTHOR LIST

Rafael Martín Nieto	rafael.martinn@uam.es
Fulgencio Navarro Fajardo	fulgencio.navarro@uam.es

CHANGE LOG

Version	Data	Editor	Description
0.1	29-05-2014	Rafael Martín Nieto	Initial version
0.2	16-06-2014	José M. Martínez	Revision
1.0	16-06-2014	José M. Martínez	First version

CONTENTS

1. Introduction

In this document, we describe different tracking research lines. The first works describe different tracking systems based on different techniques that solve some of the problems in video object tracking. Also, some video object tracker fusion methods based on combining the output of different algorithms has been developed showing improvement over individual trackers from state of the art. Finally a correlation study between some of the state of the art object tracking evaluation metrics is presented, showing redundancy between them.

All these works have been developed and implemented by the Video Processing and Understanding Lab in the Escuela Politécnica Superior of the Universidad Autónoma de Madrid.

1.1. Document structure

This document contains the following chapters:

- Chapter 1: Introduction to this document
- Chapter 2: Describes a real-time long-term single target object tracker.
- Chapter 3: Describes a feature based tracker.
- Chapter 4: Presents an improvement on a state of the art tracker.
- Chapter 5: Presents a fusion system which improves the performance of several object trackers.
- Chapter 6: Shows a correlation study of a set of video object trackers evaluation metrics.

2. Video object tracking based on template refinement of point matching

This section describes a complete framework for real-time long-term single target object tracking [1]. The obtained system supports high appearance changes in the tracked objects, occlusions, and it is also capable of recovering objects lost during the tracking process. This tracking system has been designed for automatic teacher tracking and recording of virtual lectures. Additionally, multiple sequences with real lectures and challenges have been recorded and published in order to test the designed system[1].

A two phases system has been designed for the proposed single-target object tracker. The first phase uses the PKLT to track fast displacements and to choose the object features (using color and motion coherence). The second phase (refinement) uses mean shift gradient descent to take the bounding box to the exact position of the object model, using object features provided by the PKLT learning model. Figure 2 shows the block diagram of the system.

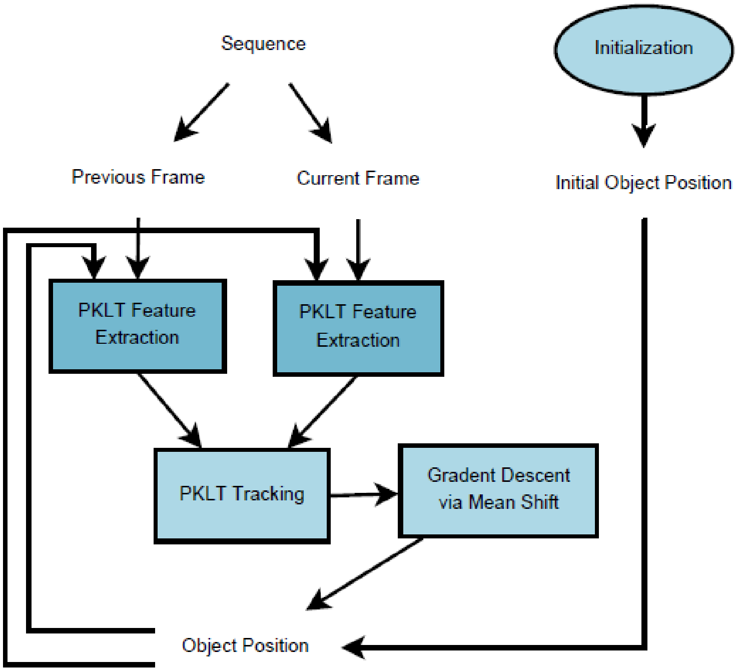

Figure 1. Block diagram of the system.

Only the position of the object in the first frame of the video is given to the initialization block. This position can be obtained in different ways, depending on the target application: it can be selected manually or by using an automatic object detector able to get the initial position

[1] http://www-vpu.eps.uam.es/publications/TeacherTrackingForAutomaticLecturesProduction/

of an object. The initial position does not need to be a bounding box, but it is possible to start from any closed polygon. In the experiments of this work, the initial object position is selected manually drawing a rectangle in the torso of the tracked person.

The KLT feature tracker is originally based on the work done by Lucas-Kanade for calculating the optical flow, subsequently completed by Tomasi-Kanade, and finally presented and clarified by Shi-Tomasi. This technique is based on characteristic points tracking, using the equations developed by Lucas-Kanade for calculating the optical flow, which also implements the iterative Newton-Raphson method for searching gradient descent. Starting with the characteristic points obtained by the method of Shi-Tomasi this tracking method calculates for each point its displacement vector with respect to the time. Taking each characteristic point as a center, a window is defined on which the gradient descent is performed using the method developed by Tomasi-Kanade.

For each frame, the method to decide the preserved points is as follows: KLT points are calculated in the current frame and the point matching (between the previous and the current frame) is performed with the KLT minimizing error process. After this, the descriptor of the points of the new frame is generated. Weights are given using the point matching which are used to discard points with low confidence. The lower weights are assigned to those points with a match that does not fit the global motion of the object.

The experiments show that the system is able to track a target during a lecture overcoming the challenges and difficulties that it might face. The main contribution of this research is a system which uses a relatively simple but well founded combination of existing methods in the state of the art capable of tracking an object (designed for tracking people but, since it works with points, generalizable to object) during a lecture and able to recover it once it has been lost during the tracking.

3. Tracker guided with SP-SIFT fine-tunned by histogram matching

This study presents an application of a recently proposed feature, SP-SIF [2]. Even if it appeared as an evolution of SIFT, which is not frequently used for tracking applications, its capacity of handling occlusion and clutter and maintaining the original SIFT properties, leads in a really appropriate feature for tracking.

For its application, a state of the art tracking main strategies analysis was made, and the results were the following:

a) Direct matching: Problems handling occlusions and target self-changes as size, shape and appearance.

b) Discriminative classification: Problems with clutter and again with target self-changes.

Evaluating the tracking strategies main problems, we reached the conclusion that features used in those strategies could be the cause and also the solution to some of them, especially in the direct matching strategy.

Analyzing in more detail the features used in tracking, we built a table comparing their behaviors against some of the most common challenges in tracking, see Table 1.

Features	Size	Shape	Appearance	Motion	Occlusion	Clutter	Box Accuracy
POI	+++	+++	++	++	+	+	+
Region	+	++	++	++	++	+++	++
Template	+	+	+	+++	++	++	+++

Table 1: Features behavior against tracking challenges. +++More robust, +Less robust.

Target self-changes mainly affect to template based methods. POI-based methods are not the best strategy to deal with occlusions and clutter due the POI description processes. Region shows an average behavior in most of the tasks.

As we did for the previously used features, we evaluate our feature against the same challenges, and the obtained results were similar to previous POI approaches in most of the challenges, except for occlusions and cluttering handling, were SP-SIFT overcame the previous results, see Table 4.

Features	Size	Shape	Appearance	Motion	Occlusion	Clutter	Box Accuracy
SP-SIFT	+++	+++	++	++	+++	+++	+

Table 2: Features behavior against tracking challenges. +++More robust, +Less robust.

With these properties in mind, we designed a direct matching tracker, based in SP-SIFT features. The main idea of the proposed method is a frame by frame SP-SIFT detection and description, and a matching with a SP-SIFT model of the target.

This initial idea presented good results in general, not losing the target and being able of handling almost every kind of situation. However, despite the great properties of the proposed features, we realized that, the lack of box accuracy showed in the challenges evaluation for the features, leads in a poor precision in the tracked element position estimation, and the results were not comparable to the state of the art tracking results.

To overcome this situation, a refinement post SP-SIFT position estimation was included. To this aim, a simple Mean Shift algorithm was included. The final scheme were the following, see Figure 1.

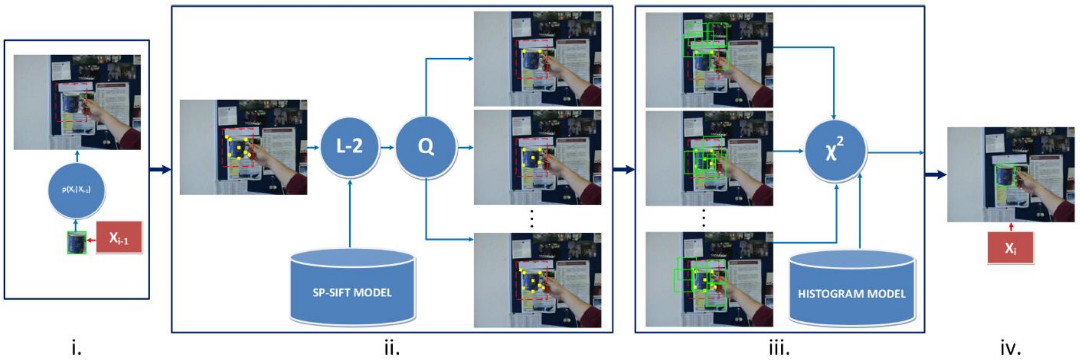

Figure 2: Tracker process for a generic frame i. i) Searching area prediction from the result of frame X_{i-1} ii) SP-SIFT stage: detection, matching (L2-norm) and quantification of matching distances iii) MEAN SHIFT refinement stage: h_i plausible states and χ^2 distance best candidate finding iv) X_i as the result for frame i

After an exhaustive evaluation, the conclusions were: average results indicate that the use of the SP-SIFT feature increases over 15% performance respect to using SIFT, which validates our hypothesis. Results respect to SoA algorithms also show that a quite simple combination of a point-based technique with a template-matching beats most of the compared approaches.

4. Superpixel Tracking based on Superpixel-SIFT

This section presents a study [3] which results in a contribution to a state of the art tracking method. The original method is called Superpixel Tracking (SPT), and has been top-ranked in recent state of the art surveys. It is presented as robust combination of two of the most promising techniques in the computer vision: discriminative strategies for tracking, and superpixels for image segmentation.

The core of the SPT algorithm lies on a segmentation on superpixels of the target searching area and its description with HSI histograms. Those descriptions are classified in foreground and background by a comparison with a previously generated HIS-histogram model. A confidence map arises from this comparison, and the final decision is taken after a *maximum a posteriori* (MAP). Avoiding occlusion algorithms are also included in the method.

The main weakness of the method is the non-discriminative technique used for the descriptions of the superpixels. So many other algorithms have been proven to be much more discriminative and effective for matching than the HSI histograms, especially the POI based. A previously proposed technique for including POI descriptions in superpixels is included in the original SPT in other to overcome its weaknesses.

Looking at the obtained results, see Table 3, a 12% of improve is obtained by our method against the results of the showed by the SPT in the original paper. The other rows are evaluations of other state of the art trackers. In the light of these results, a top-ranked tracker has been developed with the inclusion of our proposed feature, in the core of an existing method.

	MS	PF	IVT	Frag	MIL	PROST	VTD	SPT	**SPT SP-SIFT**
lemming	236	184	14	84	14	23	98	7	***3***
liquor	137	28	296	31	165	22	155	9	***5***
singer	116	25	5	21	20	-	***3***	4	***5***
basketball	203	21	120	14	104	-	11	***6***	***10***
woman	32	76	133	112	120	-	109	9	***8***
transformer	46	46	131	47	33	-	43	13	***12***
bolt	204	34	386	100	380	-	14	6	***4***
bird1	330	137	230	228	270	-	251	***15***	***15***
bird2	73	75	115	24	13	-	46	***11***	***12***
girl	304	16	184	106	55	-	57	21	***15***
average	**168,1**	**64,2**	**161,4**	**76,7**	**117,4**	**22,5**	**78,7**	**10,1**	***8,9***

Table 3. Number of successful frames. Evaluation metric of the PASCAL VOT object detection

5. Evaluation of Bounding Box Level Fusion of Single Target Video Object Trackers

The main objective of this work[4][5] is to evaluate a simple fusion system which improves the performance of several object trackers, within a methodological and rigorous evaluation framework. The considered algorithms are mono-camera single target trackers. The sequences selected in this evaluation try to represent different real scenes and conditions.

There are multiple tracking algorithms in the state of the art. The trackers used in this work are: Template Matching (TM), Mean-Shift (MS), Particle Filter-based Colour tracking (PFC), Lucas-Kanade tracking (LK), Incremental learning for robust Visual Tracking (IVT), Tracking Learning Detection (TLD), Corrected Background Weighted Histogram tracker (CBWH) and Scale and Orientation Adaptive Mean-Shift Tracking (SOAMST).

The implemented fusion methods have been selected based on its simplicity and independence (i.e., only using its outputs - bounding boxes). Fusion considered uses only the resulting bounding box of each of the individual trackers. For each frame, the bounding box resulting from the processing of each single tracking algorithm is extracted, and then the corresponding fusion is performed. As only the resulting bounding box from each tracker is used for the fusion, no matter what kind of single tracker is used for the fusion and any tracker can be used for these types of fusion. As the cost of fusing the outputs of each tracker is significantly lower than the cost of processing a frame for each independent tracker, the final cost would be similar to the cost of the slower independent tracker if they are properly parallelized.

With the (simple) fusions performed, more generalized tracking approaches have been obtained, which are able to function reasonably well in most situations (covered by the selected dataset), overcoming the problem of specialization observed in the individual trackers.

6. Correlation study of video object trackers evaluation metrics

This section presents a correlation study[6] of a set of video object trackers evaluation metrics. There are multiple metrics in the state of the art, and the main differences between them are based on the penalties that are attributed to the errors (false positives, false negatives, target loss, ...). This study was performed using multiple tracking algorithms, and an extensive set of video sequences that attempt to cover many different situations. The correlation obtained between pairs of metrics is shown in Table 4.

	SFDA	ATA	ATEi	AUCi	Overlap	CTM	TC	CoTPSi
SFDA	1,00	0,99	0,64	0,96	0,96	0,99	0,89	0,88
ATA	0,99	1,00	0,64	0,96	0,96	0,99	0,89	0,88
ATEi	0,64	0,64	1,00	0,60	0,60	0,64	0,61	0,42
AUCi	0,96	0,96	0,60	1,00	0,99	0,96	0,89	0,92
Overlap	0,96	0,96	0,60	0,99	1,00	0,96	0,89	0,92
CTM	0,99	0,99	0,64	0,96	0,96	1,00	0,89	0,88
TC	0,89	0,89	0,61	0,89	0,89	0,89	1,00	0,78
CoTPSi	0,88	0,88	0,42	0,92	0,92	0,88	0,78	1,00

Table 4. Correlation between metrics

A similar correlation (around 0.9) is shown in most comparisons between any two metrics. There is only one exception with the ATEi metric. As explained in its definition, the ATE metric can be seen as a false positive rate. This means that this measure, unlike the rest, does not penalize the existence of false negatives, as it only considers the number of false positives. In general, the metrics are highly correlated, as all attempt to measure how well the target object is tracked. The main differences between the metrics are based on the penalties that are attributed to the errors.

Just one single metric can be chosen to extract general conclusions without losing precision. When choosing a metric, the ATE must be chosen if the false positives are the only thing to consider. In the remaining cases, a single metric provides enough information. The choice depends on the specific application and other factors.

7. Conclusions and future work

This document has described different tracking research lines. The three tracking methods show promising results when compared to state of the art trackers, using different techniques in each of the systems. A common aspect of all is that they use point features, a technique that is being used in the trackers currently developed by the scientific community due to the comparatively demonstrated better results.

The trackers fusion method also shows interesting conclusions. By combining different independent methods you can obtain better results that best result given by each of them separately.

Additionally, the video object tracker metrics correlation study presents interesting results that help in deciding which method to choose for the evaluation of any tracker.

There are many possibilities for future work. One of them is to develop extensions and improvements on the different tracking system: scale change support, initialization background correction to correct the generated model in the first frame, features improvement, computational improvements, etc.

References

[1] Antonio Gónzalez Huete: “Seguimiento y producción automática mediante cámaras PTZ en entornos de red”, Master Thesis, Escuela Politécnica de Madrid, Universidad Autónoma de Madrid, Sep. 2013

[2] Fulgencio Navarro, Marcos Escudero-Viñolo, Jesús Bescós: “SP-SIFT: enhancing SIFT discrimination via super-pixel-based foreground–background segregation”, IET Electronics Letters, 50(4):272-274, Feb. 2014

[3] Fulgencio Navarro, Marcos Escudero-Viñolo, Jesús Bescós: “Enhancing region-based object tracking with the SP-SIFT feature”, in Proc. of CBMI 2014.

[4] Rafael Martín, José M. Martínez: “Evaluation of Bounding Box Level Fusion of Single Target Video Object Trackers”, in Proc. of HAIS 2014.

[5] Rafael Martín, José M. Martínez: On the fusion of single-target video objects tracking algorithms, Master Thesis, Escuela Politécnica de Madrid, Universidad Autónoma de Madrid, Sep. 2013

[6] Rafael Martín, José M. Martínez: "Correlation study of video object trackers evaluation metrics," IET Electronics Letters , 50(5):361-363, Feb. 2014