MVA2007 IAPR Conference on Machine Vision Applications, May 16-18, 2007, Tokyo, JAPAN

A Review of Tracking Methods under Occlusions

Zui Zhang
Faculty of IT, UTS
Broadway, NSW, Australia
zuiz@it.uts.edu.au

Massimo Piccardi
Faculty of IT, UTS
Broadway, NSW, Australia
massimo@it.uts.edu.au

Abstract

Object tracking in computer vision refers to the task of tracking individual moving objects accurately from one frame to another in an image sequence. Several tracking methods have been proposed in the recent literature capable of coping with a certain degree of occlusions of the objects. However, no comparative analysis of such methods has been presented to date and both the expert and the newcomer to this area may be confused about the relative effectiveness of each method when compared under the same level of complexity of the dynamic scene. In order to fulfill this need, this paper proposes a set of analysis criteria and provides a comparative review of the main recent tracking methods, in particular with respect to their capability of tracking objects under occlusions.

1. Introduction

Object tracking in computer vision refers to the task of automatically tracking individual moving objects accurately across successive frames in an image sequence. Object tracking is an essential and probably the most difficult task in visual surveillance as occlusions will interfere with the object tracking process, thus increasing the tracking difficulty. Occlusions occur when the view of a moving object is blocked completely or partially by other objects such as the static scene's background or other moving objects.

Several tracking methods have been proposed in the literature. The main, non-mutually exclusive categories that we identify are: region-based tracking, active contour-based tracking, feature-based tracking, model-based tracking and body part-based tracking. Region-based tracking algorithms track objects by identifying a connected region in the image - a "blob"- associated with each object and then tracking it over time using a cross-correlation measure. Active contour-based tracking methods track objects by representing their outlines as bounding contours and updating such contours dynamically in successive frames. Feature-based tracking methods perform recognition and tracking of objects by extracting elements such as corners and vertices, clustering them into higher level features and then matching the features between images. Model-based tracking refers to any tracking methods using models for the objects. Such models typically include shape, motion and appearance features. Priors are commonly used, such as ellipsoids for an object's shape and constraints on motion modes. At each frame, the frame pixels are matched against the existing object models, and models updated. Body part-based tracking algorithms relate specifically to people tracking. They track by detecting and tracking body parts and then re-assembling them together to detect the complete human body. For all such methods, occlusions pose a significant challenge as they occur repeatedly and extensively in real scenes, affecting the accuracy of tracking. In recent years, several methods have started to explicitly deal with object occlusions. However, no comparative analysis of such methods has been presented to date to compare the effectiveness of each method under the same level of complexity of the dynamic scene. As the number of proposals keeps on growing, the lack of understood performance becomes more problematic. In order to overcome this limitation, this paper proposes a set of analysis criteria and provides a comparative review of the main recent (2002-2006) tracking methods based on their capability and effectiveness in tracking objects under occlusions.

The rest of the paper is organized as follows: Section 2 describes the main features of each reviewed method. Section 3 proposes the comparison criteria and the assessment rationale for tracking under occlusion. Section 4 presents the comparison of these reviewed methods based on the criteria introduced in Section 3. Conclusive remarks are addressed at the end of this paper.

2. The Reviewed Approaches

2.1. Tao, Sawhney and Kumar (2002)

In [1], Tao, Sawhney and Kumar have proposed a method for tracking objects under occlusions by capturing the spatial and the temporal constraints on the shape, motion and appearance of the tracking objects in a dynamic layer representation. Each object, including the background of the scene, is defined as a "layer" and modeled by four features: appearance, approximate elliptical shape, position and motion. Such features are not allowed to freely vary between successive frames. Rather, they have to undergo a "coherency model" where smooth, gradual changes are assumed to be more likely. This is simply obtained by using a constancy model added with a noise component for each feature. At every new frame, the optimization of the layer representation is achieved by applying an Expectation-Maximization (EM) algorithm to re-estimate the membership of pixels to layers and the layers' parameters with maximum a-posteriori probability (MAP). As it is not computationally practical to address this optimization simultaneously for all features, the EM algorithm in [1] uses four separate formulas for the layer ownership and the motion, shape and appearance models, respectively. As only gradual changes can be accommodated, short-term occlusions do not significantly affect the objects' models.

2.2. Zhou and Tao (2003)

Authors from [8] have extended and improved the method proposed in [1] by using multiple occluding background layers and explicitly inferring the depth ordering of foreground layers. Occluding background layers are regions of the static scene occurring to be in front of some moving objects. Depth ordering is the relative depth of the foreground layers from the closest to camera to the furthest to the camera. By adding those two concepts in the model, the prior function of the dynamic layer representation now becomes the product of the depth ordering, the foreground layer shape prior, the background layer shape prior, motion model and foreground layer appearance.

2.3. Wu and Nevatia (2005, 2006)

In [2], Wu and Nevatia have proposed a two-step method which is a significant advancement upon other methods recognising a person as a whole. This method takes into account the deformable nature of humans as objects. The first step for this method is to use part detectors to detect body parts based on a set of edgelet features. The second step is to combine the results from various part detectors and formulate the detection of multiple people, possibly occluding each other, using MAP on a joint likelihood model. The articulated human model is very simple, with three body parts: head & shoulders, torso and legs. The main feature is the *"*edgelet", defined as the line segment of an edge of a body part. There are three steps involved in detecting body parts: the first is to build a weak classifier for each edgelet feature. Then, the AdaBoost algorithm is applied to linearly combine sets of weak classifiers into strong classifiers. Finally, nesting structured detectors are constructed on the output of the strong classifiers. The joint likelihood consists of two parts, the state of the occluded person and the responses from the part detectors. Based on the assumptions and responses from the head detector and the full-body detector, an occupancy map and visibility for each occluded person's body parts are built. The match between the responses and parts are done in a greedy way based on a distance matrix until no more valid pair is available.

In [3], the authors extend their part detector proposed in [2] to a full tracking method for multiple, partially occluded humans. The new idea is to use a forward human tracking algorithm which combines the static human detection with data association & meanshift tracking to achieve tracking of multiple, partially occluded humans by body parts. The responses from the static part detector is taken as input into the human tracker and then matched against the object hypotheses. This match is done through a greedy algorithm similar to that used for part detection. The human tracking process is done using two strategies, data association and meanshift tracking. If a new response matches a current hypothesis, then the hypothesis grows based on data association; otherwise, meanshift tracking is used to re-assess correspondences. The basic idea for the meanshift tracking is to track by using a probability distribution combining an appearance model, a dynamic model, and the detection confidence.

2.4. Zhao and Nevatia (2003, 2004b)

In crowded scenes, single foreground blobs often contain multiple people due to substantial occlusions. In [4], a method has been proposed for segmenting single humans in such blobs by using a Markov Chain Monte Carlo (MCMC) algorithm. MCMC methods are a class of algorithms for sampling from probability distributions based on constructing a Markov Chain that has the desired distribution as its stationary distribution. The model dynamics consist of two components, namely jump and diffusion. The jump component can describe changes in the number of people in the blob, while the diffusion component describes the variations in the values of the parameters. In the method, a 3D human model is used consisting of four ellipsoids corresponding to head, torso and two legs. Each ellipsoid is controlled by four parameters, namely length, "fatness", position and orientation. Three model types for movements are considered: both legs together, left leg forward, and right leg forward. Head hypotheses are formulated in two different ways depending on whether the heads lie at the boundary of the foreground blob or its interior.

In [6], the foreground segmentation method of [4] is extended to become a full tracking method. The extension also includes a revision of several aspects of the previous method, such as an improved prior function, re-formatting the parameters of the human model, and adding a likelihood at the single person level. Such a single-person likelihood favours both the difference to the background and the similarity with its correspondence in the previous frame, which enables simultaneous detection and tracking of an object. This is achieved by iteratively optimizing the joint likelihood of all single-person likelihoods. The search space is limited by computing the joint likelihood at each iteration incrementally in the neighborhood of the object which is being changed. The state corresponding to the maximum posterior value up to the current iteration is recorded and becomes the final solution when the given number of iterations is reached. The Kalman filter is used to filter and predict the states of each object.

2.5. Zhao and Nevatia (2004a)

Another tracking method explicitly devised to track people has been proposed in [5]. It uses both a top-view and a side-view of the dynamic scene to generate a 3D human shape model for identifying and tracking each person. Constructing the human model and segmenting each occluded person are performed in several steps. The first is to allocate all the head tops in the frame. The next is to find the approximate height for the person. Once the head top and the height are determined, an ellipsoid human model can be generated to represent the human on the ground plane. The process of identifying each human figure in a frame is a recursive process as some people may not be identified in the first round as they are occluded by other people. The iterative process involves segmenting human bodies out of the foreground, removing the identified humans and their shadows from the foreground mask, and performing an opening on the foreground mask to remove isolated residuals.

The tracking process involves generating two

templates that are used by a Kalman filter to predict the position for each person in the current frame and compute the best match between the predicted human model and an extracted body within a search region. As people are matched in turn starting from the nearest to the camera, the relative space is marked as occupied, thus reducing the search region for people not yet matched.

2.6. Zhou, Chellappa, and Moghaddam (2004)

In [7], a particle filter is used for tracking objects in the presence of occlusions. The particle filter is based on a simple, linear state transition equation, but the velocity and the noise statistics are adapted along the time, thus making the model more flexible. The number of particles in the filter is also adapted along the time. The features used for tracking consist only of appearance features: observations consist of 2D image patches in grey levels; the appearance model of each object is pixel-based and consists of a mixture of Gaussians at each pixel. The appearance model is kept up to date by an EM algorithm that is invoked upon every new matching observation.

An interesting aspect of the method in [7] is in its explicit detection and handling of occlusions. Whenever a pixel in the image patch is too distant from the corresponding mixture-of-Gaussian model, it is labeled as an outlier. Then, if the number of outliers in the image patch is above an assigned fraction, an occlusion is declared and the appearance model and velocity are not updated. Such an approach can effectively prevent incorrect updating of the model even for extended periods of time. However, it makes the method prone to false detection of occlusions in case of sudden illumination changes.

3. Analysis Criteria

In the following, we present the set of analysis criteria for the comparative review of the methods in Section 2:

Model space. The number of dimensions used to model the human object and the scene; either 2D or 3D.

Features. The features used to track or match target objects, according to categories shape, colours and motion.

Shape model. The model used to represent the shape feature of the tracked object. The model is divided into contours and basic shapes. Contours can be edgelets or B-splines. Basic shapes are an approximation of the object's global shape and include circles, rectangular bounding boxes, ellipses and ellipsoids and their combinations.

Motion model. The motion of tracked objects, that can be modeled either as rigid or non-rigid depending on the object category. Normally, vehicles are modeled as rigid and humans are modeled as non-rigid. In addition, motion can be assumed as constant or varying through time.

Occlusion handling. The degree of occlusions that the tracking algorithm is capable of handling. The most critical aspects of occlusions are their extent in space and duration in time. We decided to categorise these aspects into coarse levels as in the following. We have also considered the nature of the occlusions and the possible changes in an object's appearance occurring during occlusions due, for instance, to illumination variations.

1. ***Fraction*** is the fraction of the tracking object being occluded by another layer.
   partial(≤ 0.3), significant(≤ 0.5), large(>0.5)
2. ***Duration*** is the time duration of the occlusion as a proportion to the total tracking time. This could also be defined using the number of frames. In both cases, standard frame rate is assumed.
   short(≤ 1s or 5 frames), long(>1s or 5 frames)
3. ***Layers*** is the type of object/layer for which occlusion can be identified and handled.
   background, foreground
4. ***Degree of change*** is the degree of change in appearance for an occluded object during the occlusion.
   negligible(≤ 20%), partial(≤ 50%), large(>50%)

An obvious way to assess the reviewed methods against occlusions could be that of testing them on a common benchmark. However, not only this would prove highly time consuming, but it would also be very difficult to tune parameters as could be done by their authors. Instead, we decided to use the following indicators to measure the methods' capabilities against occlusions: 1) assessing the footage presented in the respective papers against the above criteria. It is in the interest of the authors that occlusion cases displayed in figures be representative of probing situations that the method can properly handle; 2) the authors' own statements; 3) intrinsic properties of the approaches chosen.

Appearance model. The appearance of the tracking object. It can be modeled in various ways, from simple colour component histograms at the object level, up to pixel-level models modelling correlations of each pixel in both spatial and temporal dimensions.

Statistical methods. Statistical methods are the core components of tracking approaches. Examples of the most popular statistical methods used in the reviewed approaches are listed as follows, ranking from the least to the most time consuming: Kalman filter, EM, Particle filter, Hidden Markov model, Meanshift. Ranking is necessarily approximate as the actual computational complexity depends on a number of factors, not least how many parameters in the statistical method require statistical estimation at their turn.

4. Performance Analysis

The following comparative tables report the performance comparison between the reviewed six methods in accordance with the analysis criteria explained in the previous section. From Tables 1-3, several conclusions regarding the benefits and drawbacks of each method can be obtained.

Table 1. Comparing the methods: features & statistics

Method (section)	Model space	Features			Statistical methods
		Shape	Colours	Motion	
2.1	2D	*	*	*	EM
2.2	2D	*	*	*	EM

2.3	2D	*	*	N/A	Kalman, Meanshift, AdaBoost
2.4	3D	*	*	N/A	Kalman, MCMC
2.5	3D	*	*	N/A	Kalman
2.6	2D	N/A	*	N/A	Particle

Kalman: Kalman filter; Particle: Particle filter

Table 2. Comparing the methods: models

Method (section)	Shape model		Appearance model	Motion model
	Contour	Basic shape		
2.1	N/A	Ellipsis	Colour Corr	R, CV
2.2	N/A	Ellipsis	Colour Corr	R, CV
2.3	Edgelets	Ellipsis	Colour Hist	N/A
2.4	N/A	Ellipsoid	Colour Hist	N/A
2.5	N/A	Ellipsoid	Colour Hist	R, CV
2.6	N/A	N/A	Colour Corr	AV

R: rigid motion; CV: constant velocity; AV: adaptive velocity

Table 3. Comparing the methods: occlusion handling

Method (section)	Occlusion handling			
	Fraction	Duration	Layers	Degree
2.1	Partial	Short	Fg	N/A
2.2	Signif	Long	Fg, Bg	N/A
2.3	Signif	Long	Fg, Bg	N/A
2.4	Partial	Short	Fg, Bg	N/A
2.5	Partial	Short	Fg, Bg	N/A
2.6	Signif	Short	Fg	N/A

Signif: Significant; Fg: foreground; Bg: background

A significant strength of the method in Section 2.1 is in its use of a broad range of features. This provides the potential to survive occlusions since features may not degrade simultaneously to the same extent. However, a main limitation seems in the fact that the method does not model occlusions explicitly. Furthermore, the 2D model chosen for the objects' shape seems suitable only for a limited number of scenarios such as aerial cameras or planar motion. The additions reported in Section 2.2 extend the method to deal with background occlusions and mutual occlusions between targets and make it generally much more resilient to occlusions.

The method in Section 2.3 represents a significant advancement towards effective tracking of humans in crowded scenes. Table 1 clearly shows that its major disadvantage lies in the high computational cost deriving from the tracking of individual body parts and the use of meanshift tracking. On the other hand, with the continuous increase in processors' speed, the use of sophisticated statistics will soon be compatible with real-time constraints.

The method in Section 2.4 is the only one designed to cope with merged blobs from multiple humans. However, it may not prove robust enough to handle human tracking when the occlusions are severe.

The method presented in Section 2.5 combines two views of the same scene in order to retrieve depth order of multiple targets, thus promising to be effective in handling mutual occlusions. Other aspects omitted in the review, such as detection of shadows and mirror images, make it very robust to real-life situations. On the other hand, adequate views may not be available in existing camera networks, thus requiring infrastructure changes.

Eventually, the method at 2.6 has a major strength in the continuous adaptation of the parameters in the state transition and observation equations of its particle filter. Furthermore, it explicitly detects major occlusions so as to prevent "pollution" of the object's model. However, it is prone to falsely recognise sudden illumination changes as occlusions. This would either limit the duration of the occlusions handled, or cause missed tracking of targets.

5. Conclusions

In this paper, we have presented a comparative review of tracking methods under occlusions. Such a review aims to provide the reader with a rapid and thorough understanding of these methods, in particular with respect to their capability of handling occlusions in both space and time. Among the six reviewed methods, significant difference in performance are reported. Methods such as [1] do not explicitly model occlusions, but still prove robust to them to a certain extent. More sophisticated methods such as [2-8] adopt specific solutions to the occlusion problem. However, no method yet seems to have tackled certain aspects such as, for instance, appearance changes occurring during occlusions. It is also desirable that the best features from the various reviewed methods be re-unified in a combined approach.

References

[1] Tao, H., Sawhney, H. S. & Kumar, R. 2002, 'Object Tracking with Bayesian Estimation of Dynamic Layer Representations', *IEEE Transactions on Pattern Anal. and Machine Intell.*, vol. 24, no. 1, pp. 75-89

[2] Wu, B. & Nevatia, R. 2005, 'Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detectors', *Proc. of IEEE Int. Conf. on Computer Vision*, vol. 1, pp. 90-97

[3] Wu, B. & Nevatia, R. 2006, 'Tracking of Multiple, Partially Occluded Humans based on Static Body Part Detection', *Proc. of Comp. Vis. and Patt. Recogn.*, vol. 1, pp. 951-958

[4] Zhao, T. & Nevatia, R. 2003, 'Bayesian human segmentation in crowded situations', *Proc. of Computer Vision and Pattern Recognition*, vol. 2, pp. 459-466

[5] Zhao, T. & Nevatia, R. 2004a, 'Tracking Multiple Humans in Complex Situations', *IEEE Transactions on Pattern Anal. and Machine Intell.*, vol. 26, no. 9, pp. 1208-1221

[6] Zhao, T. & Nevatia, R. 2004b, 'Tracking multiple humans in crowded environment', *Proc. of Computer Vision and Pattern Recognition*, vol. 2, pp. 406-413

[7] Zhou, S. K., Chellappa, R., & Moghaddam. B., 2004, 'Visual tracking and recognition using appearance-adaptive models in particle filters', *IEEE Transactions on Image Processing*, vol. 13, no. 11, pp. 1491-1506

[8] Zhou, Y. & Tao, H. 2003, 'A background layer model for object tracking through occlusion', *Proc. of the Ninth IEEE Int. Conf. on Computer Vision*, vol. 2, pp. 1079-1085