Tracking based on local motion estimation of spatio-temporally weighted salient points

Vincent Garcia*, Éric Debreuve, and Michel Barlaud
Laboratoire I3S, Université de Nice - Sophia Antipolis, France
{garciav,debreuve,barlaud}@i3s.unice.fr

Abstract

The extraction of a video object contour, called "rotoscoping" in cinematographic post-production, is usually performed manually and frame by frame. Semi-automatic algorithms have been proposed to reduce the load of this task. However, they classically use region information and are usually based on a notion of homogeneity of the object. This homogeneity description might be difficult to establish and, consequently, the tracking may be not precise enough. The proposed method relies on the analysis of some temporal trajectories of salient points, or keypoints, called tracks. The main contribution of this paper is the local estimation, both spatially and temporally, of the contour motion from these tracks. The proposed method seems accurate, robust to outliers, and allows local deformation. Moreover, it can deal with partial occlusions.

1. Introduction

The segmentation of video objects is a low level task required for many applications, for example in cinematography. The term "rotoscoping" used in cinematographic post-production corresponds to the all-digital process of tracing outlines over digital film images to produce digital contours in order to allow special visual effects. The segmentation is usually performed manually and frame by frame by so-called animators. As a consequence, it is a long, repetitive, and expensive task. This is why rotoscoping is a very active research topic of video processing. The rotoscoping problem is too complex to define a fully-automatic algorithm. In this paper, we focus on the tracking of an object (*i.e.*, the extraction of the object contour for all frames of the sequence) given an initial, hand-edited contour in the first frame.
Some tracking methods [4, 10] perform object tracking using a bounding box. These methods are more adapted to scene analysis and understanding. Methods using global (*i.e.*, region) information [1, 5] are usually based on a notion of (possibly non-trivial) homogeneity of the object (*e.g.*, intensity, motion, histogram...). If the object is complex or has a complex motion, this homogeneity description might be difficult to establish and not precise enough to guarantee an accurate tracking.
The proposed method is based on temporal trajectories of keypoints called *tracks*. The contribution of this paper is the local approach, both spatially and temporally, used to estimate the local contour motion from the tracks. A spatial weighting allows to track precisely object with local deformations. A temporal weighting applied in a group of pictures (GOP) allows to account for more observations for the parameter estimation process. This improves the precision of the estimation and the robustness to outliers compared to an estimation using only the next frame. The proposed method seems accurate, robust to outliers, and allows local deformations. Moreover, it can deal with partial occlusions.
The paper is organized as follows: Section 2 focuses on the proposed method. Section 3 shows and discusses some results using a measure of quality. Finally, Section 4 concludes.

2. Proposed approach

Let V be a video composed of n frames $F^1, \cdots, F^n$. Let C^1 be a hand-edited contour in frame F^1 segmenting the object of interest. The contour is discretized in a set of samples according to its representation (polygon, spline, *etc.*). The tracking problem consists in defining the contours $C^2, \cdots, C^n$ from C^1 in order to segment the object in the successive frames. The process has to be precise because the occurrence of an error for the computation of a contour C^i can induce an important drift of the following contours.

2.1. Building of tracks

A *track* is the temporal trajectory of a *keypoint*. A keypoint [7], also called *interest point*, *salient point*, *feature*

*This work was partly supported by *"Le Conseil Régional Provence-Alpes-Côte d'Azur"*, France.

point in the literature, is a point in an image which has a well-defined position and can be robustly detected, for example object corners. Combined with local descriptors [9], they are distinctive and have proven to be useful in different applications [2]. Keypoints extracted from different frames are matched based on their descriptors, typically using the L1-norm or L2-norm. SIFT [9] is a popular descriptor. However, the Harris keypoint extractor combined with a gray-level circular window descriptor is appropriate for our application. Tracks are built as follows. First, keypoints are extracted independently in each frame of the video using a chosen keypoint extractor. Then, the descriptor is computed for each keypoint and each keypoint is matched with keypoints defined in the next frame by matching their descriptors (using the L1-norm) with cross-validation. The motion between two consecutive frames is assumed to be small enough to consider candidate keypoints only in a search window. Finally, pairs of matching keypoints sharing a common keypoint are concatenated into sets of keypoints called *tracks* (see Figure 1). Tracks are generally not defined for all the frames of the video but on a subset of frames. The following notation will be used: a track T_k defined on the interval $[F^i, F^j]$ is a set of keypoints $T_k = \{t_k^i, ..., t_k^j\}$.

Figure 1. Tracks on sequence Erik.

2.2. Motion estimation

2.2.1 Basic procedure

The proposed tracking method is based on the following assumption: the overall motion of the object contour can be deduced from the motion of the keypoints belonging to the object. In other words, the object contour is guided over time by the tracks remaining inside the pipe formed by the hand-edited contours C^1 in frame F^1 and all the contours C^i computed so far.

In the following, we assume that the object contours $C^1, ..., C^m$ already exist. The contour C^j, defined on the frame F^j, is a set of samples $\{p_1^j, \cdots, p_l^j\}$. The problem is to compute C^{m+1} from the previous contour C^m and the tracks. First, each track T_k remaining inside the temporal pipe formed by $C^1, ..., C^m$ and defined at least for the frames F^m and F^{m+1} is selected. Second, the pairs of keypoints $\{t_k^m, t_k^{m+1}\}$ are extracted from selected tracks. Third, an affine motion matrix M is estimated from these pairs using the following M-estimator [6]:

$$M = \arg\min_M \sum_k f(\|M.t_k^m - t_k^{m+1}\|), \qquad (1)$$

where t_k^m and t_k^{m+1} are given in projective coordinates, M is a 3×3 affine motion matrix, f a cost function, and $\|.\|$ stands for the Euclidean norm. The minimization is performed by using a simplex method [8].

We have chosen M-estimators because they provide a precise estimation of the parameters and because they are robust to outliers (presence of keypoints of the background in the contour) as opposed to the classical mean square error corresponding to choosing $f(x) = x^2$. In this paper, we chose f equal to the absolute value. Finally, the contour C^{m+1} is deduced by applying M to the samples of C^m:

$$p_i^{m+1} = M.p_i^m, \ \forall i \in [1, l], \qquad (2)$$

where p_i^m and p_i^{m+1} are given in projective coordinates.

2.2.2 Spatio-temporal approach

The approach proposed in Section 2.2.1 allows global affine deformation of the contour. However, local deformation cannot be estimated. To overcome this issue, we propose to compute a local motion from the same set of selected tracks. Parameter estimation requires enough observations to be robust and accurate. With small objects, the number of selected tracks may be as low as 10 tracks. Consequently, only 10 observations are used for the parameter estimation. Increasing of the number of tracks, by increasing the sensitivity of the keypoint detector to extract more keypoints, would decrease their relevance and consequently the accuracy of the parameter estimation. Instead, we propose in addition to extract pairs of keypoints on previous and next frames from the selected tracks.

Spatially local motion estimation – Let p_i^m be a sample of the contour C^m. The local motion is computed by estimating for each sample p_i^m an affine motion matrix M_i. Then, p_i^{m+1} is deduced by applying M_i to p_i^m. The idea is to give more influence to tracks spatially close to p_i^m. A weighting $\lambda_{i,k}$, function of a distance between keypoint t_k^m and the sample p_i^m, is associated with each pair $\{t_k^m, t_k^{m+1}\}$:

$$\lambda_{i,k} = \varphi(\|p_i^m - t_k^m\|), \qquad (3)$$

where φ is positive, monotonically decreasing function defined on $\mathbb{R}^+$. For example, φ may be a Gaussian function

or a function derived from the classical regularization functions. The Euclidean distance is chosen for $\|.\|$. The motion matrix M_i is estimated by minimizing the following weighted M-estimator [6]:

$$M_i = \arg\min_M \sum_k \lambda_{i,k}.f(\|M.t_k^m - t_k^{m+1}\|), \quad (4)$$

Finally, p_i^{m+1} is deduced by applying M_i to p_i^m as in (2).

Both spatially and temporally local motion estimation – We make the following assumption: in a video, it is reasonable to assume that, within a group of picture (GOP), the motion of points (keypoints and samples) is stationary (or conversely, the size of GOP must be chosen such that this assumption is reasonable). Let us consider that the motion is stationary in a GOP of G frames. G should be chosen even so that the GOP can be centered around frames $\{F^m, F^{m+1}\}$. Let g be equal to $\frac{G}{2}$. Therefore, at most G pairs of keypoints are extracted for each selected track:

$$\{t_k^{m-g}, t_k^{m-g+1}\}, \{t_k^{m-g+1}, t_k^{m-g+2}\}, \cdots$$
$$\cdots, \{t_k^m, t_k^{m+1}\}, \cdots, \{t_k^j, t_k^{j+1}\}, \cdots, \{t_k^{m+g}, t_k^{m+g+1}\}$$

It is implicitly assumed that the tracks remaining inside $C^1, \cdots, C^m$ will remain inside $C^{m+1}, \cdots, C^n$. Given a set of keypoint pairs, the main idea is to give more importance to pairs spatially close to p_i^m and temporally close to F^m. The temporal weighting δ_j for a pair of keypoints $\{t_k^j, t_k^{j+1}\}$ is given by:

$$\delta_j = \phi(|m - j|), \quad (5)$$

where ϕ has the same properties of φ. Motion matrix M_i is estimated as follows:

$$M_i = \arg\min_M \sum_k \lambda_{i,k}. \sum_{j=m-g}^{m+g} \delta_j.f(\|M.t_k^j - t_k^{j+1}\|). \quad (6)$$

The spatial weighting $\lambda_{i,k}$ is computed for a track T_k and then is equal for all the pairs extracted from this track. On the contrary, the temporal weightings δ_j are different for each pair extracted from T_k but similar for all k. The functions φ and ϕ use the same weighting function but their input values are differently stretched to be defined on a common interval. Among tested functions, the Gaussian provides the best results.

3. Experiments

3.1. Sequences without occlusions

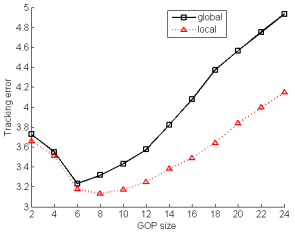
The visual measure does not allow to quantify the error done in comparison to the real object contour manually defined. Given a frame F^i (for i a frame index in $[1, n]$), let C_g^i ("g" for ground truth) be the true object contour manually defined, and C_c^i the computed object contour. We define an error measurement representing a mean distance between C_g^i and C_c^i $\forall i$. Let A_g^i be the mask computed from C_g^i, and A_c^i be the mask computed from C_c^i. The pixels of the masks are equal to 1 inside the contour and 0 outside. The contour error for the frame F^i, given in percentage of misclassified pixels, is the following:

$$d^i = 100 \times \frac{\sum_{x,y} A_g^i(x,y) \oplus A_c^i(x,y)}{\sum_{x,y} A_g^i(x,y)} \quad (7)$$

where $\oplus$ is the xor operation and (x, y) the pixel coordinates. The tracking error $\overline{d}$ for the whole sequence is the average contour error for all computed contours.

The proposed method has been tested on natural CIF

(a) Hand-edited contour C^1.



(b) Tracking error.

Figure 2. Tracking results on sequence Carmap.

video sequence (CIF=320×240 pixels) of 36 frames named *Carmap* (see Figure 2(a)) where the object contour is composed of 4 samples. Some local deformations appear due to the viewpoint displacement. Figure 2(b) illustrates the tracking error as a function of the GOP size for the spatially global and the local approach. The spatially global approach consists in computing a local motion estimation using a similar spatial weighting for each keypoint pair extracted (*i.e*, $\lambda_{i,k} = 1\ \forall i, k$). The local approach is the method proposed in Section 2.2.2. The spatially local approach allows to decrease the tracking error. Also, the tracking error has a minimum. This minimum, depending on the weighting function used, is reached for the optimal GOP size for which the assumption of motion stationary is verified. It varies depending on the tolerance on outliers. Below this optimum, there is not enough data for a reliable estimation. Beyond this optimal value, the assumption appears wrong and consequently the tracking error increases. Therefore, according to Figure 2(b) and in comparison to the basic global approach (see Section 2.2.1), the local approach, both spatially and temporally, decreases the tracking error. More generally, the local approach provides a better tracking than the global approach.

3.2. Occlusion management

The proposed method allows to manage partial object occlusion as can seen on the SD (SD=720 $\times$ 480 pixels) sequence *Driving* of 30 frames (see Figure 3) where the object contour is composed of 29 samples. In case of occlusion, for example on frame F^{m+1}, a part of the object of interest is hidden by an other object denoted by O. Some tracks of object O remains to C^{m+1} but do not remains to $C^1, \cdots, C^m$. They are not selected by the first step of the proposed method, and, they are consequently not used for the parameter estimation allowing to manage occlusions. The ability of the proposed method to track occluded object does not depend on the proportion of object occluded. Indeed, if there is at least one track remaining to $C^1, \cdots, C^{m+1}$, the method is still able to track the object.

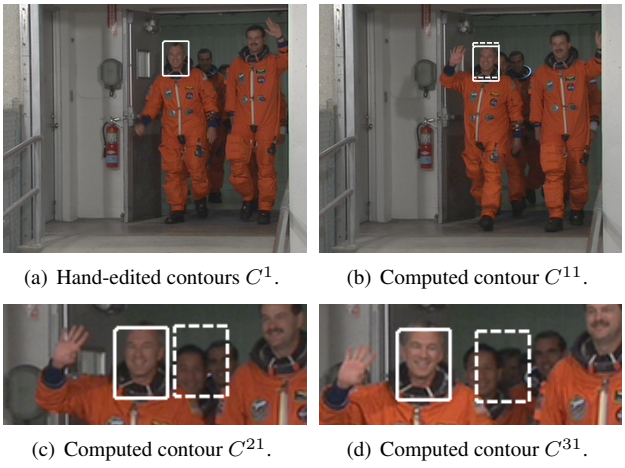
Figure 3. Hand-edited contour C^1 and computed contour C^{28} on sequence Driving.

3.3. Application to bounding box tracking

The proposed method allows to track precisely an object contour. Used with a bounding box as initial contour and a spatially global motion estimation, the proposed method has an application to tracking based on object bounding box. Here, the proposed method is compared with a mean-shift based method [4] (an implementation is found in [3]). The two methods are tested on the CIF sequence *Crew* of 80 frames where the initial bounding box is initialized around the head of an astronaut. Figure 4 shows the resulting trackings. On this sequence, the proposed modified method provides an accurate tracking throughout the sequence while the mean-shift based method fails.

4. Conclusion

We have proposed a tracking method relying on the analysis of some temporal trajectories of salient points, or keypoints, called *tracks*. The main contribution of this paper is the local estimation, both spatially and temporally, of the contour motion from these tracks. The experiments show that the local approach provides a more precise tracking than with a global motion estimation. The proposed method seems accurate, robust to outliers, and allows local deformation. Moreover, it can deal with partial occlusions.



(a) Hand-edited contours C^1. (b) Computed contour C^{11}.

(c) Computed contour C^{21}. (d) Computed contour C^{31}.

Figure 4. Computed contours with proposed method (plain line) and mean-shift based method (dashed line) on sequence Crew.

References

[1] A. Blake and M.Isard. *Active Contours: The Application of Techniques from Graphics,Vision,Control Theory and Statistics to Visual Tracking of Shapes in Motion*. Springer–Verlag, New-York, tats-Unis, 1998.

[2] S. Bres and J. Jolion. Detection of interest points for image indexation. In *Int. Conf. Visual Information and Information Systems*, pages 427–434, 1999.

[3] R. Collins, X. Zhou, and S. K. Teh. An open source tracking testbed and evaluation web site. In *IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance*, January 2005.

[4] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *Computer Vision and Pattern Recognition*, pages 142–151, Hilton Head Island, South Carolina, 2000.

[5] É. Debreuve, M. Gastaud, M. Barlaud, and G. Aubert. A region-based joint motion computation and segmentation on a set of frames. In *Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*.

[6] C. Gourieroux and A. Monfort. *Statistics and Econometric Models*, volume 1. Cambridge University Press, 1995.

[7] C. Harris and M. Stephens. A combined corner and edge detector. In *Proc. of The Fourth Alvey Vision Conference*, pages 147–151, Manchester, UK, 1988.

[8] J. C. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright. Convergence properties of the nelder-mead simplex algorithm in low dimensions. *SIAM Journal on Optimization*, 9:112–147, 1998.

[9] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. Journal of Computer Vision*, 60(2):91–110, 2004.

[10] R. Venkatesh Babu, P. Pérez, and P. Bouthemy. Robust tracking with motion estimation and kernel-based color modelling. In *Proc. Int. Conf. on Image Processing (ICIP'05)*, Genova, Italy, September 2005.