REGION-OF-INTEREST TRACKING BASED ON KEYPOINT TRAJECTORIES ON A GROUP OF PICTURES

Vincent Garcia, Éric Debreuve, and Michel Barlaud

Laboratoire I3S, Université de Nice - Sophia Antipolis
2000 route des Lucioles, 06903 Sophia Antipolis, FRANCE
{garciav,debreuve,barlaud}@i3s.unice.fr

ABSTRACT

This paper deals with region-of-interest (ROI) tracking for applications such as video surveillance or cinematographic post-production. An ROI is typically delineated by a bounding box or a basic shape such as an ellipse in the first frame of the video. The tracking problem consists in detecting the ROI throughout the video as it moves and deforms. This detection can be done based on the full content of the ROI. However, since the ROI is, by definition, an approximate segmentation of the actual object of interest, it includes some background. This can make the ROI detection less accurate and induce a drift. Instead, we propose to use keypoint extractors and local descriptors combined with robust motion estimation. The motion estimation relies on the analysis of the temporal trajectories, or *tracks*, of the keypoints in groups of pictures (GOP). Some results are presented on natural sequences. The proposed method seems accurate.

1. INTRODUCTION

Object tracking is a challenging task for the research vision community. It is a low level task required for many different applications as video surveillance, television, or cinematographic post-production. The shape of the contour used to track the object depends on the type of application. Indeed, a basic bounding box is acceptable for a video surveillance application while Bezier curves are used for tasks requiring precision such as compositing. In this paper, we focus on ROI tracking where the ROI is defined on the first frame of the sequence by a hand-edited shape.

Some tracking methods based on active contours [1, 2] propose to track the object of interest by computing the object contour frame by frame. However, they are usually based on a notion of (possibly non-trivial) homogeneity of the object (*e.g.*, intensity, motion, histogram...). This homogeneity description might be difficult to establish. ROI-based methods, like bounding box tracking [3, 4], are an interesting alternative when the exact object contour is not required. However, when using global ROI descriptors, they may not be accurate enough if the object appearance changes too much.

The proposed tracking approach is based on *keypoints* [5, 6, 7]. Keypoints (also called *interest points*, *salient points*, or *feature points* in the literature) have proved their usefulness for image indexation [8], image retrieval [9], and 3D object model tracking [10, 11, 12]. The proposed method is based on temporal trajectories of keypoints called *tracks*. The main contribution of this paper is the estimation of the object motion performed from the tracks on groups of pictures (GOP). The use of GOP allows to increase the reliability of the parameter estimation. According to the results, the proposed method seems accurate.

The paper is organized as follows: Section 2 focuses on the proposed tracking method. Section 3 shows and discusses some tracking examples. Finally, Section 4 concludes.

2. PROPOSED APPROACH

Let V be a video composed of n frames $F^1, \cdots, F^n$. Let B^1 be the contour of the ROI in frame F^1. The tracking problem consists in defining the contours $B^2, \cdots, B^n$ from B^1. The process has to be precise because an error in the computation of a ROI B^i can induce an important drift of the following ROI.

The proposed tracking is based on the analysis of some reliable temporal trajectories called *tracks*. We assume that the overall motion of "large" objects can be estimated from the trajectory of so-called keypoints.

2.1. Building of tracks

A *track* is the temporal trajectory of a *keypoint*. A keypoint [5, 6, 7], also called *interest point*, *salient point*, *feature 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 [13, 14, 15, 16, 17], they are distinctive and have proven to be

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

useful in different applications [8, 9, 10, 11, 12, 18]. Keypoints extracted from different frames are matched based on their descriptors, typically using the L1-norm or L2-norm. Descriptors have different properties. For example, a window of gray-level or color values of an image is spatially distinctive but is not robust to rotation. On the contrary, the gray-level distribution estimated by the Parzen method [19] on a window is a descriptor robust to rotation and small scaling but is not spatially distinctive. The SIFT method [15] improves matching by defining a multi-scale keypoint extractor and a descriptor based on the gradient direction histogram. This descriptor is both spatially distinctive and robust to rotations and scalings. In this paper, we will use the Harris [5] keypoint extractor combined with a gray-level circular window descriptor. This choice is discussed in Section 2.3.

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 Fig. 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 $T_k = \{t_k^i, ..., t_k^j\}$. Tracks are the base of the proposed method and their usefulness will be justify in Section 2.2.

Fig. 1. Example of tracks on sequence *Erik* on frame F^1. Note that the tracks on the background project to a single point since the background is still. The character turns his head to his lefts.

2.2. Motion estimation of the ROI

In this Section, we introduce first a basic approach of motion estimation from tracks. Second, we propose to improve the tracking by using groups of pictures (GOP).

2.2.1. Basic procedure

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

In the following, we assume that the object ROI $B^1, ..., B^m$ already exist. The ROI B^j, defined on the frame F^j, is a set of samples $\{p_1^j, \cdots, p_l^j\}$. The problem is to compute B^{m+1} from the previous ROI B^m and the tracks. First, each track T_k remaining inside the temporal pipe formed by $B^1, ..., B^m$ 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 [20]:

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

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 [21].

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 ROI) as opposed to the classical mean square error corresponding to choosing $f(x) = x^2$. Function f can be chosen among [22, 23]:

$$f(x) = |x| \tag{2}$$
$$f(x) = \sqrt{x^2 + \epsilon^2} - \epsilon \tag{3}$$
$$f(x) = 2\log(\cosh(x)) \tag{4}$$
$$f(x) = \log(1 + x^2) \tag{5}$$
$$f(x) = \frac{x^2}{1 + x^2} \tag{6}$$

In this paper, we chose f equal to the absolute value. Finally, the contour B^{m+1} is deduced by applying M to the samples of B^m:

$$p_i^{m+1} = M.p_i^m, \ \forall i \in [1, l], \tag{7}$$

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

2.2.2. Temporally local motion estimation

M-estimators, as most of statistics methods, require enough observations to provide a robust and precise estimation of the parameters. 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 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. Thus, we propose to account for more keypoints on previous and next frames from the selected tracks. 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 ROI $B^1, \cdots, B^m$ will remain inside $B^{m+1}, \cdots, B^n$. Given a set of keypoint pairs, the main idea is to give more importance to pairs 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:

$$\delta_j = \psi(|m - j|), \tag{8}$$

where ψ is positive, monotonically decreasing function defined on $\mathbb{R}^+$. For example, ψ may be a Gaussian function or a function differentiated from the classical regularization functions [23]. The motion matrix M is estimated using the following weighted M-estimator [20]:

$$M = \arg\min_M \sum_k \sum_{j=m-g}^{m+g} \delta_j . f(\|M.t_k^j - t_k^{j+1}\|), \tag{9}$$

where t_k^j and t_k^{j+1} are given in projective coordinates, M is a 3×3 affine motion matrix, $f(x) = |x|$, and $\|.\|$ stands for the Euclidean norm. The temporal weightings δ_j are different for each pair extracted from T_k but similar for all k. Finally, the ROI B^{m+1} is deduced by applying M to each samples of B^m as in Eq. (7).

2.3. Discussion on choice of keypoint extractors and descriptors

The keypoint extraction [5, 6, 7] and the definition of local descriptors [13, 14, 15, 16, 17] have been intensively studied. Indeed, the keypoint matching have been used in many application like image and video indexing and retrieval [8, 9, 24] or 3D object model tracking [12, 10, 11]. The recent descriptors, for example SIFT [15], allow to define reliable keypoint matching between different views of an object, particularly in case of important viewpoint displacement. In our case, the viewpoint displacement and the motion of the objects between two consecutive frames appears small. The matching problem performed on a search window is easier. The matching performed using the Harris extractor combined with a gray-level circular window as a descriptor provides keypoint matches as reliable as using a reference method as SIFT. As a consequence, this choice of keypoint extractor and descriptor does not appear critical. For this reason, we have chosen to use the Harris keypoint extractor combined with a gray-level circular window descriptor. The radius chosen for the descriptor is 8 pixels.

3. EXPERIMENTS

3.1. Comparison of the proposed method to two other approaches

In this section, the proposed method is compared with two other tracking methods. The first one is a simple block matching method [25] using a sub-optimal approach [26]. The object ROI hand-edited in frame F^1 is the initial block. The blocks corresponding to the optimum of a given similarity measure are detected on frames $F^2, \cdots, F^n$. We chose sum of absolute differences as the similarity measure. The second method is a mean-shift based method [3, 27]. Only the visual quality of the resulting trackings was compared. For the experiments, the size of the GOP was 4 frames. The contours computed with the three tested tracking methods are presented superimposed over the frame. The ROI was a rectangle with corners $\{p_1^j, p_2^j, \cdots, p_4^j\}$.
First, the three methods were compared on the SD (SD=704×576 pixels) sequence *Ice* of 10 frames (see Fig. 2). The rectangle is initialized for all methods on frame F^1 around the head of a skater (see Fig. 2(a)). The three methods provide an efficient tracking. However, the proposed method and the block matching based method seem to track better the object than the mean-shift based method.
Second, the methods were compared on the CIF (CIF=352×288 pixels) sequence *Crew* of 80 frames (see Fig. 3). The rectangle is initialized for all methods on frame F^1 around the head of an astronaut (see Fig. 3(a)). Up to frame 50, both the block matching based method and the proposed method provide an accurate tracking (see Fig. 3(b)) while mean-shift based method fails. After frame F^{50}, both the block matching based method and the mean-shift based method fail to track the head (see Fig. 3(c) and Fig. 3(d)).
Third, the methods were compared on the CIF sequence *Football* of 20 frames. The rectangle is initialized for all

methods on frame F^1 around the helmet of the football player (see Fig. 4(a)). With this sequence, only the proposed method provides an accurate tracking (see Fig. 4). The two other methods are unable to track the football player helmet. The proposed method seems robust to outliers (presence of keypoints of the background in the ROI), and luminance change (see Fig. 3).

3.2. Tracking using an elliptic shape

An elliptic shape seems more adapted than rectangle bounding boxes for face tracking. Indeed, during the step of track selection, the use of bounding box may select tracks which do not belong to the object. The keypoint pairs extracted from these tracks appears as outliers. Although the use of robust motion estimation allows to decrease their influence, some applications, like face colorization, typically use ellipses to track objects.
The rectangle bounding box of the previous examples has been replaced with an ellipse. 50 samples $\{p_1^j, p_2^j, \cdots, p_{50}^j\}$ were used.
Fig. 5 shows the tracking performed on sequence *Football* presented in Section 3.1.

4. CONCLUSION

We have proposed a tracking method based on the estimation of the region-of-interest (ROI) motion from keypoint temporal trajectories. The method is independent of the shape of the ROI. The use of groups of pictures increases the number of observations and thus increases the precision and the robustness to outliers of the motion estimation. The proposed method performs well on several natural sequences.

5. 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. 1

[2] 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)*. 1

[3] 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. 1, 3

[4] R. Venkatesh Babu, P. Pérez, and P. Bouthemy. Robust tracking with motion estimation and kernel-based color modelling. In *IEEE Proc. Int. Conf. on Image Processing*, Genova, Italy, September 2005. 1

[5] 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. 1, 2, 3

[6] S. M. Smith and J. M. Brady. Susan – a new approach to low level image processing. *Int. Journal of Computer Vision*, pages 45–78, may 1997. 1, 3

[7] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *Int. Journal on Computer Vision*, 60(1):63–86, 2004. 1, 3

[8] S. Bres and JM. Jolion. Detection of interest points for image indexation. In *Int. Conf. Visual Information and Information Systems*, pages 427–434, 1999. 1, 2, 3

[9] Ch. Wolf, J. M. Jolion, W. Kropatsch, and H. Bischof. Content based image retrieval using interest points and texture features. In *IEEE Int. Conf. on Pattern Recognition*, 2000. 1, 2, 3

[10] M. Armstrong and A. Zisserman. Robust object tracking. In *Proc. Asian Conf. on Computer Vision*, pages 58–61, 1995. 1, 2, 3

[11] E. Rosten and T. Drummond. Fusing points and lines for high performance tracking. In *IEEE Int. Conf. on Computer Vision*, Beijing,China, 2005. 1, 2, 3

[12] D. G. Lowe. Robust model-based motion tracking through the integration of search and estimation. *Int. Journal of Computer Vision*, 1992. 1, 2, 3

[13] Timor Kadir and Michael Brady. Saliency, scale and image description. *Int. Journal of Computer Vision*, 45(2):83–105, 2001. 1, 3

[14] Y. Ke and R. Sukthankar. Pca-sift: A more distinctive representation for local image descriptors. In *IEEE Proc. on Computer Vision and Pattern Recognition*, 2004. 1, 3

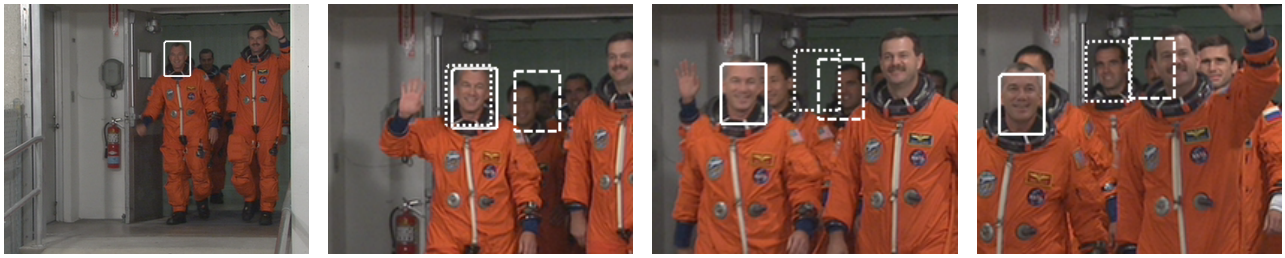
[15] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. Journal of Computer Vision*, 60(2):91–110, 2004. 1, 2, 3

[16] F. Tonnin, P. Gros, and C. Guillemot. Analysis of multiresolution representations for compression and local description of images. In *Int. Conf. on Visual Information and Information Systems*, pages 234–246, July 2005. 1, 3

[17] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005. 1, 3

(a) Initialization of the object bounding box on frame F^1 (b) Close up of the computed bounding boxes on frame F^4 (c) Close up of the computed bounding boxes on frame F^7 (d) Close up of the computed bounding boxes on frame F^{10}

Fig. 2. Comparison of tracking results on 10 frames of sequence *Ice* for the block matching based method (dotted line), the mean shift based method (dashed line), and the proposed method (solid line). The head of the skater is well tracked over the 10 frames with all three methods.



(a) Initialization of object bounding box on frame F^1 (b) Close up of the computed bounding boxes on frame F^{30} (c) Close up of the computed bounding boxes on frame F^{60} (d) Close up of the computed bounding boxes on frame F^{80}

(e) Thumbnails showing the ROI computed with the proposed method and for every 4 frames from frame F^1 to F^{69}.

Fig. 3. Comparison of tracking results on 80 frames of sequence *Crew* for the block matching based method (dotted line), the mean shift based method (dashed line), and the proposed method (solid line). The head of the astronaut is precisely tracked over the 80 frames using the proposed method. On the contrary, the tracking fails after 50 frames using the block-matching based and mean-shift based methods.

[18] M. Brown and D. G. Lowe. Recognising panoramas. In *IEEE Int. Conf. on Computer Vision*, page 1218, Washington, DC, USA, 2003. IEEE Computer Society. 2

[19] E. Parzen. On the estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33:1065–1076, 1962. 2

[20] C. Gourieroux and A. Monfort. *Statistics and Econometric Models*, volume 1. Cambridge University Press, 1995. 2, 3

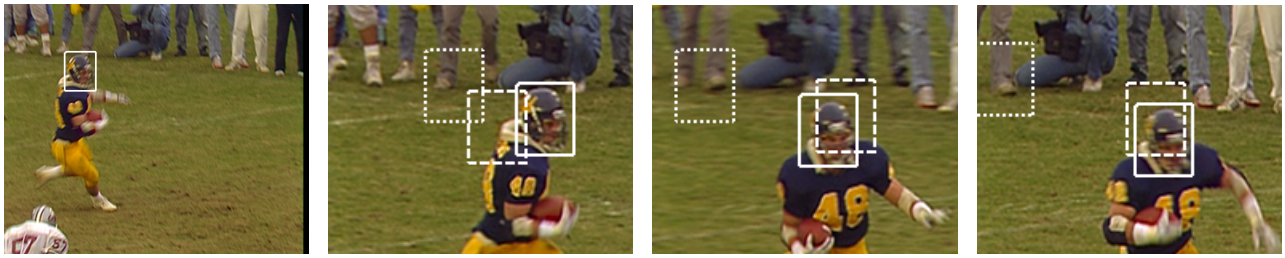
[21] J. C. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright. Convergence properties of the nelder-mead simplex algorithm in low dimensions. *SIAM Journal on Optimization*, 9:112–147, 1998. 2

[22] M. J. Black and P. Anandan. The robust estimation of multiple motions: parametric and piecewise-smooth flow fields. *Comput. Vis. Image Underst.*, 63(1):75–104, 1996. 2

[23] P. Charbonnier, L. Blanc-Feraud, G. Aubert, and M. Barlaud. Deterministic edge-preserving regularization in computed imaging. *IEEE Trans. in Image Processing*, 6(2):298–311, February 1997. 2, 3

[24] F. Souvannavong, B. Mérialdo, and B. Huet. Region-based video content indexing and retrieval. In *Int. Workshop on Content-Based Multimedia Indexing*, June 2005. 3

[25] A. Hirano Y. Iijima T. Koga, K. Linuma and T. Ishiguro. Motion compensated interframe coding for video conferencing. *Proc. Nat. Telecommun. Conf.*, 1981. 3



(a) Initialization of object bounding box on frame F^1 (b) Close up of the computed bounding boxes on frame F^7 (c) Close up of the computed bounding boxes on frame F^{14} (d) Close up of the computed bounding boxes on frame F^{20}

(e) Thumbnails showing the bounding box computed with the proposed method and for every frames from frame F^1 to F^{18}

Fig. 4. Comparison of tracking results on 20 frames of sequence *Football* for the block matching based method (dotted line), the mean shift based method (dashed line), and the proposed method (solid line). The head of the football player is precisely tracked over the 20 frames using the proposed method. On the contrary, the tracking fails after 3 frame using the block-matching based and mean-shift based methods.

(a) Initialization of object ellipse on frame F^1 (b) Computed ellipse on frame F^7 (c) Computed ellipse on frame F^{14} (d) Computed ellipse on frame F^{20}

Fig. 5. Tracking results on 20 frames of sequence *Football*. The head of the football player, detected by an ellipse, is precisely tracked over the 20 frames of the sequence.

[26] S. Zhu and K.K. Ma. A new diamond search algorithm for fast block-matching motion estimation. *IEEE Trans. On Image Processing*, 9(2), February 2000. 3

[27] Robert Collins, Xuhui Zhou, and Seng Keat Teh. An open source tracking testbed and evaluation web site. In *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS 2005), January 2005*, January 2005. 3