Orientation histogram-based matching for region tracking

David Marimon and Touradj Ebrahimi
Signal Processing Institute (ITS)
Ecole Polytechnique Fédérale de Lausanne (EPFL)
CH-1015 Lausanne, Switzerland

Abstract

A region tracking technique with particular emphasis on rotation robustness is presented. It is based on region matching divided in two consecutive steps, gradient orientation histogram matching and template matching through Normalised Cross Correlation (NCC). Given the orientation histograms of two image patches, a novel technique is used to estimate the rotation between them together with the similarity. This estimation enhances the performance and speeds-up the process of patch recognition. Fast computation of histograms using the integral histogram approach [9] is exploited. Experiments show a high accuracy in the estimation of location and orientation.

1. Introduction

Region tracking is the task that deals with the localisation and trailing of an image patch or region, inside a video sequence. This task is often used for object tracking [2] with applications such as video surveillance [1], robot localisation [11] and face detection [13], to mention a few. Similar to this task is feature or key point tracking [6] with applications such as augmented reality [8], pose estimation [5] or mosaicing, among others. The difference comes from the type of saliency. While key point tracking refers to locating a single salient point in the scene (either 3D or 2D), usually detected in a previous extraction phase, for region tracking it is a region in space with a certain intensity or texture characteristic that is sought. This paper concentrates on this latter type of tracking. The difficulties that visual tracking has to face are mainly viewpoint and illumination changes, which in some applications have to be addressed in real time.

In this paper, we present a novel non-trained fast orientation histogram-based matching technique with application to region tracking. The matching is divided in two consecutive steps: gradient orientation histogram matching followed by template matching on the pixel neighbourhood of the best histogram matches. The process is initiated by first exhaustively searching a patch p inside a video frame. A similarity map is built by comparing the gradient orientation histogram extracted from the neighbourhood around each point in the image with the one extracted from the patch p. For this comparison, we define a novel metric called *Circular Normalised Euclidean Distance* (CNED) that is used to obtain the most similar circularly shifted histogram, together with the degree of similarity itself. The second step is a Normalised Cross Correlation (NCC) that takes into account the rotation estimated (provided by the CNED). The position of the most correlated point inside the video frame is considered the centre of the patch that is being tracked. The correlation coefficient (result of the NCC) at this point indicates the quality of the track. With the proposed technique it is also possible to provide a rough estimate of the patch rotation in the current frame.

This paper is structured as follows. Related research is presented in the next section. The orientation histogram-based matching technique is explained in Section 3. It describes the histogram descriptor of the region, the metric used to compare histograms and the final template matching step. Experiments are dealt with in Section 4 followed by the conclusion in Section 5.

2. Related works

Both region and feature point tracking techniques can be classified into two categories: statistical tracking and tracking by detection. *Statistical tracking* is mostly based on motion prediction/update filters. These typically use the Kalman filter, multimodel filters [12] or, more generally, Monte Carlo approaches (particle filters) [4]. *Tracking by detection* localises regions or points, despite their previous location, in a frame-by-frame basis. These techniques commonly use classification schemes or invariant descriptors. For the first group, classifiers are trained with a set of positive and negative patch examples. Research is concentrated on this training set and the classification structure used, and generally focuses on real-time applications [5, 13]. For the second group, attention is paid to the description of the region or keypoint neighbourhood [7]. In this case, more invariance to viewpoint (or to affine transformations) and illumination changes is often achieved, counterbalanced with higher computational complexity at run time when compared to the classification-based techniques [1, 2, 6, 11].

Among these techniques, some works have identified the potential of using histogram-based descriptors for visual tracking [2] and, more recently in [1, 11]. Comaniciu [2] presented an object tracking framework where the contribution of pixels to the histogram are weighted with a convex and monotonic decreasing kernel profile. The advantage of this kernel is that the effect of peripheral pixels is lessened. It uses different information for the histogram depending on the application and the Bhattacharyva distance for histogram matching. In many cases, histograms have their information concentrated in a little portion of bins. Both Adam [1] and Serratosa [11] have chosen the Earth's Movers Distance (EMD) [10] which handles this situation and is specially tailored for intensity histograms, dealing with illumination changes. Adam [1] presented a fragments-based framework which also deals with partial occlusions. The problem of viewpoint changes is not addressed directly by none of these works. On the other hand, Lowe [6] uses the spatial distribution of gradient orientation histograms for viewpoint and illumination invariance.

In this work, stable scale-space points are extracted and described for further matching with very impressive results. However, the application of this technique to visual tracking is limited by the extraction phase. If the object or region to track is small, it cannot be known a priori if it will be represented by any key point.

Our technique is situated in between the two groups of tracking by detection techniques. On the one hand, several positive examples are used. On the other hand, no classifier is used, but instead, a strong descriptor focused on rotation and illumination invariance is built.

3. Orientation-based Matching

The matching technique presented here takes two steps that provide discrimination and speed-up at the same time. These steps are, firstly, an exhaustive gradient orientation histogram matching followed by template correlation using the results of the first.

For each video frame I, the procedure runs as described next. An exhaustive search of a patch p in the image is first performed using histogram matching. At each point in I, an N-bin histogram of its neighbourhood is computed. To speed-up the computation of the histograms, we take advantage of the integral histogram approach [9]. Computing the contribution of a single bin to the histogram can be performed with only four memory accesses. With this approach, the necessary steps to compute any gradient orientation histogram in the image are as follows. Firstly, the gradient of I is computed at each point (x, y), given $dy = I(x, y+1) - I(x, y-1)$ and $dx = I(x+1, y) - I(x-1, y)$, as follows

$$\begin{aligned} m(x,y) &= \sqrt{dy^2 + dx^2} \\ \theta(x,y) &= \arctan(dy/dx), \end{aligned} \tag{1}$$

where m is the magnitude and θ is the orientation of the gradient. Secondly, θ is quantised in N bins. The running sum of each bin is computed separately. In order to compact the statistical description of the patch and to reduce the effect of noise, the contribution of each point in $\theta(x, y)$ to the corresponding bin is weighted by its magnitude $m(x, y)$ (similar to the approach in [6]). Additionally, the contribution of the central part of the patch is augmented to approximate the same effect as for the kernel-based approach of [2]. Given a fixed neighbourhood size, the histogram at any point is obtained and can then be matched to the histogram of the patch p. The similarity between both histograms is computed using the Circular Normalised Euclidean Distance described in Section 3.1. The particular robust histogram that describes p is explained in Section 3.2. Once the histogram matching is done for each point in I, a similarity map is obtained. This map is used for further template matching (final step) as described in Section 3.3.

3.1. Circular Normalised Euclidean Distance

From a theoretic point of view, the gradient has a continuous response to a continuous function. Suppose that a probability density function (PDF) is computed from the continuous orientation gradient of a perfectly circular and continuous patch p. A rotation of δ degrees of the patch is a circular shift (0^o and 360^o being the same) of the gradient orientation PDF. Now, suppose that the PDF is expressed with a histogram of N bins. In this case, a rotation of δ would change the values of its bins. In particular, when $\delta = k \cdot 360/N$ with $k \epsilon \mathbb{Z}$, the histogram would be exactly equal to a perfect shift, and the shift in bins would be equal to k. This ideal case is not completely fulfilled in reality and the rotation and further calculation of the gradient introduces changes in the shape of the histogram. However, it is possible to identify this shift if a certain degree of variation of the bins is accepted.

Following this reasoning, we introduce a novel metric to compare orientation histograms (or, generically, circular vectors), the *Circular Normalised Euclidean Distance* (CNED). Not only the CNED measures the distance d between two vectors, but it also determines the circular shift $\hat{s}$ that corresponds to the minimal distance. In other words, CNED $= [\hat{s}, d(a, b, \hat{s})]$ where $\hat{s} = \arg\min_s d(a, b, s)$,

$$d(a,b,s) = \sqrt{\sum_{i=0}^{N-1} \frac{(a(i) - b(\mathrm{rem}(i+s, N)))^2}{\sigma_a^2(i)}}, \tag{2}$$

a and b are vectors of length N, s is the shift that takes a discrete value between 0 and $N-1$, rem is the remainder function, and σ_a^2 is the variance associated to vector a.

3.2. Histogram descriptor of the patch

Texture information obtained from the gradient is chosen to generate a strong histogram descriptor of a patch. The main reason lies on the little sensibility of the gradient to illumination changes, which is one of the problems that tracking has to deal with. As described in Section 1, another major problem to tackle is viewpoint invariance. Similar to providing positive examples to a classification engine, we propose to generate several versions of the patch that is to be sought and from these versions, create a single histogram. More precisely, generate rotated versions and concentrate our efforts in a descriptor that can deal with rotations. As mentioned before, orientation histograms repeat approximately their shape every $\Delta = 360/N$ degrees (especially for large N). This can be exploited by aligning the histograms of versions rotated exactly by $k\Delta$ with $k\epsilon\mathbb{Z}$.

The steps followed to obtain the histogram descriptor are explained next. Firstly, N rotated versions of the patch p to be located are pre-computed with an angle of rotation of $n\Delta$ degrees (for $n = 0, .., N-1$) where N is the number of bins. These versions are cropped so as to eliminate additional pixels introduced by the rotation. Secondly, the histogram of each of these versions is obtained from the quantised θ weighted with m where the central part of the patch is also augmented (see the introduction of Section 3). Finally, the descriptor of the patch is the mean obtained with the N histograms aligned according to their rotation. In addition to this mean histogram, we store the related variance to enrich the description and use it in the computation of the CNED.

This average of rotated versions gives a strong descriptor when the rotation of the image is around $n\Delta$ degrees. It could be argued that for non-integer bin-wide angles higher variations will occur. However, experimentation shows that using enough bins lessens this effect and the descriptor used in conjunction with the CNED metric is reliable even around $n\Delta + \Delta/2$ degrees (see Section 4).

3.3. Template matching

The similarity map provided by the histogram matching discards many unrelated points but is not selective enough for tracking purposes. Spatial intensity information (template) is used as a further selection criterion. More precisely, those points with the most similar histograms are kept in a set S. Template matching is done using a Normalised Cross Correlation (NCC) between the templates (neighbourhood pixel intensity) around the centre points and the template of the patch p. The precision of the template comparison is increased by using the shift given by the CNED. Indeed a shift of k bins translates, as described before, into a rotation of the patch by $k\Delta$. The NCC is then computed between the neighbourhood of the point $(x, y)\epsilon S$ and the pre-computed version of p rotated by $\hat{s}(x, y) \cdot \Delta$ degrees, where $\hat{s}(x, y)$ is the shift obtained by the CNED with the histogram of the neighbourhood of (x, y).

The estimated position of the patch at the current video frame is given by the point $(x, y)\epsilon S$ with the highest NCC. The orientation is given by the corresponding shift. Although we have opted for this straightforward choice, other options are possible. For instance, the NCCs obtained for all or some of the points in S could be used as measurements to correct the state of a filter.

4. Experiments

This section presents the tracking accuracy achieved. The assessment method chosen is the comparison between the performance of the proposed technique and the ground truth obtained from several synthetic video sequences of 490 frames each. These sequences are generated with similarity transformations (rotation, translation and scaling) of a natural outdoor image (with size 300x200 pixels). Several experiments were run obtaining similar results but due to the lack of space, only the results of one of these sequences is reported here. In this sequence, an image is transformed with large rotations, 2D translations and scaling (scale factor $\in [0.8, 1.2]$). A patch is tracked and the evolution of the coordinates (in x and y axes) of its centre is compared to the ground truth. Also, the shift estimated by our matching technique is compared to the ground truth rotation. The Mean Absolute Difference (MAD) between the ground truth and the estimate of the trajectory of the patch is also provided for better comparison. Together with these results, the NCC obtained by our technique with the patch rotated according to shift (see Section 3.3) is compared to the NCC that is obtained when the patch is not rotated, both computed with the neighbourhood centered at the point given by our technique. The histogram matching step uses 20 bins ($\Delta = 18$ degrees) and the size of the patch that is being tracked is 10x10 pixels. The number of most similar histograms kept for template matching is fixed to 150.

The image used for tracking and the patch that is being searched (bounded with a square) are shown in Figure 1. This patch is obtained from the neighbourhood of a corner point detected with the Harris corner detector [3]. The tracking and Normalised Cross Correlation (NCC) results are depicted in Figures 2a and 2b, respectively.

Both good location accuracy and rotation estimation (logically limited by the number of bins) are achieved. In general, the trajectory is followed by the proposed technique, except for a few frames where an unrelated point is selected (see Figure 2a). This phenomenon is also visible in Figure 2b as a poor correlation is obtained at these frames. Such behaviour might be either because the histogram computed at the correct position is too different from the histogram descriptor and consequently this point is not among the 150 candidates, or related to a wrong orientation estimation leading to a poor correlation score.

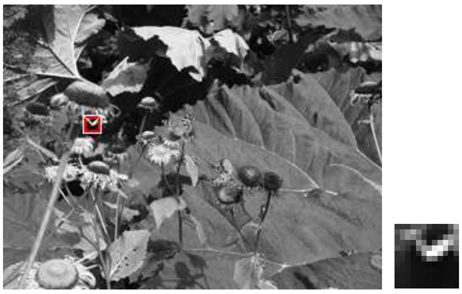


Figure 1: Image used for tracking with the patch bounded with a square (left) and zoom of the patch (right).

The mean NCC achieved by our technique for this sequence is 0.81 (NCC $\in [0, 1]$, where NCC=1 indicates complete correlation). Whereas it can be seen that the correlation without considering the orientation rapidly fails as the rotation goes beyond $\pm 10^o$. This gives an idea of the enhancement brought by our orientation centered processing. Samples of the tracked trajectory with a square bounding the estimated patch location and orientation are shown in Figure 3.

5. Conclusions

A region tracking technique has been presented. The technique is divided in two consecutive steps, gradient orientation histogram matching and template matching through Normalised Cross Correlation (NCC). A method to estimate the rotation between two image patches is exploited to enhance the template matching performance. Fast computation of histograms using the integral histogram approach [9] is exploited. Advantages of the proposed technique include the fact that there is no need to tune a classifier and the simplicity of the algorithm with a very reduced number of parameters. In addition, the regions can be selected either automatically or even manually provided that enough texture information is available.

Experiments on a synthetic video sequence shows the high accuracy and correlation achieved by the proposed technique together with a good approximation of the rotation of the region (limited by the number of bins of the histogram). In particular, the method shows robustness in front of rotation, translation and scaling with a factor $\in [0.8, 1.2]$.

Future path of research will focus on analysing the dependance of the method on the patch that is chosen, in particular, the size and available texture information. A detector of such regions providing higher tracking performance is envisioned.

6. Acknowledgements

The first author is supported by the Swiss National Science Foundation (grant number 200021-113827), and by the European Networks of Excellence K-SPACE and VISNET II. The research leading to this work was also partially in the framework of COST292 Action on Semantic Multimodal Analysis of Digital Media.

References

[1] A. Adam, E. Rivlin, and I. Shimshoni. Robust fragments-based tracking using the integral histogram. In *IEEE Conf.*

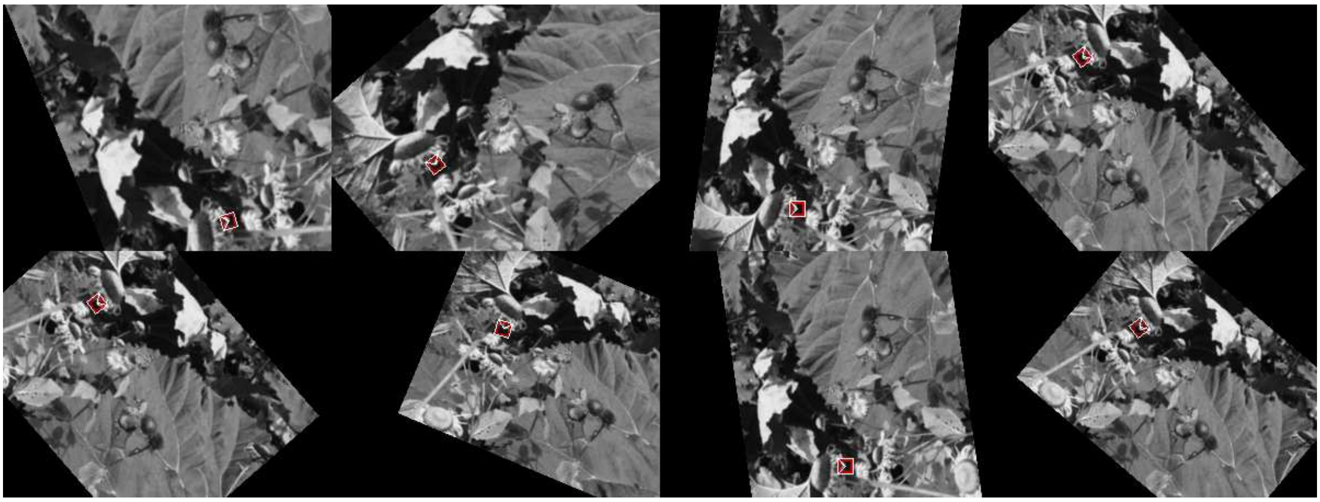


Figure 3: Tracking results sampled every 60 frames.

on Computer Vision and Pattern Recognition (CVPR), volume 1, pages 798–805, June 2006.

[2] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):564–577, 2003.

[3] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey Vision Conf.*, pages 147–151, 1988.

[4] M. Isard and A. Blake. CONDENSATION – conditional density propagation for visual tracking. *Intl. Journal of Computer Vision*, 29(1):5–28, 1998.

[5] V. Lepetit, J. Pilet, and P. Fua. Point matching as a classification problem for fast and robust object pose estimation. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 244–250, June 2004.

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

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

[8] U. Neumann and S. You. Natural feature tracking for augmented reality. *IEEE Transactions on Multimedia*, 1(1):53–64, Mar 1999.

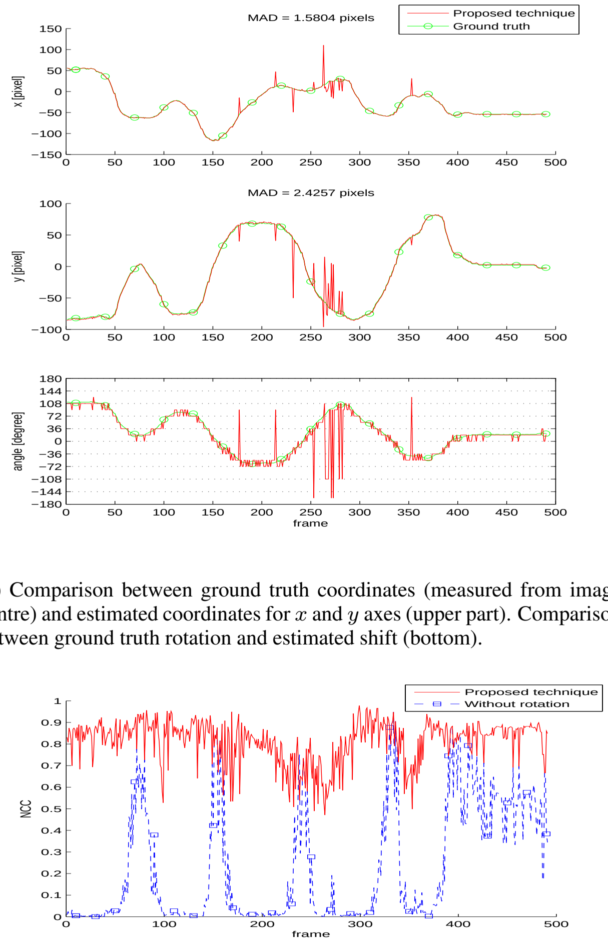
[9] F. Porikli. Integral histogram: a fast way to extract histograms in cartesian space. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 829–836, 2005.

[10] Y. Rubner, C. Tomasi, and L. Guibas. The earth mover's distance as a metric for image retrieval. *Intl. Journal of Computer Vision*, 40(2):99–121, 2000.

[11] F. Serratosa and A. Sanfeliu. Vision-based robot positioning by an exact distance between histograms. In *Intl. Conf. on Pattern Recognition (ICPR)*, pages 849–852, 2006.

[12] P. Tissainayagam and D. Suter. Visual tracking and motion determination using the IMM algorithm. In *Intl. Conf. on Pattern Recognition (ICPR)*, pages 289–291, 1998.

[13] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 511–518, 2001.



(a) Comparison between ground truth coordinates (measured from image centre) and estimated coordinates for x and y axes (upper part). Comparison between ground truth rotation and estimated shift (bottom).

(b) Comparison between NCC obtained by the proposed technique and the NCC obtained when the patch is not rotated.

Figure 2: Quantitative results for the used sequence.