

Rotation correlation maps

David Marimon and Touradj Ebrahimi
Signal Processing Institute (ITS)
Ecole Polytechnique Fédérale de Lausanne (EPFL)
CH-1015 Lausanne, Switzerland
{david.marimon, touradj.ebrahimi}@epfl.ch

Abstract

A correlation map shows the correlation between an image region and another image computed at the neighbourhood of each point of the second image. Most cross-correlation techniques fail when the viewpoint rotation grows over a certain value, usually around 20 degrees. The approach presented here aims at resolving this problem and at the same time provide an estimate of the actual rotation. This paper presents a method to obtain a correlation map robust to rotation together with an orientation map that gives an estimate of the rotation between the region and the image at each point. We call this a Rotation Correlation Map (RCM) and is built using texture and intensity information consecutively. The gradient histogram (texture distribution) of the region is used as a fast strategy to locate possible points of high correlation and an estimate of their orientation with respect to the region that is compared. Among the candidates with high histogram similarity, Normalised Cross Correlation (NCC) (using intensity) is computed using the orientation estimated in the previous step. Results show the accuracy in correlation and orientation and the overall higher performance when compared to similar maps.

Keywords Correlation, rotation, template matching, histogram matching, similarity map.

1 Introduction

Correlation is one of the most extended operations in image and video processing. A correlation map indicates the degree of correlation of an image with another image at each point of the second image. Generally, a correlation map is used to locate an image patch inside another image by detecting the maximum value in the map. Applications that use correlation maps are related to region recognition. This is the case of object tracking [1–10], and camera tracking [11–15], among others.

A correlation map can be considered a particular case of a similarity map. A similarity map, provides the similarity between an image patch and the neighbourhood centered at each point of an image. In the case of the correlation, the similarity is measured using the cross-correlation. The accuracy of a similarity map depends on several factors. Firstly, the patch that is being sought must be distinctive enough when compared with the image. For example, in grey-level images, texture information must be relevant and in colour images, the colour distribution of the patch should be non common. Secondly, the descriptor of the patch should highlight the features of that patch. Examples of features commonly described are pixel values, intensity, colour or gradient orientation distribution, the response to several filters, and mixtures of spatial and statistical information, among others. Thirdly, the measure used to compute the similarity can take advantage of these descriptors.

Most researchers have focused on descriptors and measures that are robust or even invariant to viewpoint and/or illumination distortion. This invariance leads to a loss of information. In other words, the actual viewpoint transformation between the original patch and the image is not computed, nor the exact change of illumination. In fact, the enhancement of invariance is application dependent. However, in many applications, only the localisation of the patch is used because the viewpoint transformation is not available. Hence, most of these applications could benefit of knowing for instance,

the rotation. This is the case for visual tracking applications such as those mentioned before. Hence, an advantage would be provided by a method that is capable of tackling rotations without such loss of information. In this paper, we present such a method.

In particular, we present a correlation map that provides accurate localisation of a patch inside an image. Together with this, we estimate the approximate 2D rotation that transforms the patch into that image for each point of the latter. We call this a Rotation Correlation Map (RCM) which consists in two closely related maps. One provides the estimated *rotation* of a patch in each point of the tested image. The other one, provides the *correlation* between a neighbourhood at each point of the tested image with the patch rotated according to the orientation map at that point. The RCM is built in two steps: an orientation histogram matching followed by template matching. More precisely, the gradient orientation histogram of the region is used as a fast strategy to locate possible regions of high correlation. The gradient also provides an estimate of the orientation of the region in respect to the image at each point, which gives the orientation map. Among the candidates with high histogram similarity, the NCC is computed given the rotation estimated in the previous step. The output of this step is the correlation map.

This paper is structured as follows. First, Section 2 describes research related to correlation maps. Second, the method to obtain the rotation correlation map together with a tailored region descriptor are presented in Section 3. The assessment of this method is given in Section 4. Finally, concluding remarks are dealt with in Section 6.

2 Related works

This section takes a broader look at the solutions to region recognition (or matching) in related research. Recognition techniques can be separated into two categories: trained and non-trained.

For the trained category, classifiers are trained with a test set of positive and negative patch examples. Research is concentrated on the data set and the classification techniques (e.g., [16, 17]). These techniques provide an excellent compromise between speed and performance at run-time. However, the time consumed to gather or generate the training data and train the classifier is generally high. This category is not detailed here because it is not directly related to the proposed method. The reader is referred to [18] for more details.

For the non-trained category, recognition is done by comparing the descriptor of a patch with the descriptors obtained at different locations in the image. This process can be described mathematically as follows. Given a patch (or region) \mathbf{P} and the descriptor of this patch $f(\mathbf{P})$, the similarity of \mathbf{P} with an image \mathbf{I} at point (x, y) is

$$d(f(\mathbf{P}), f(\mathbf{R}_{x,y})) \tag{1}$$

where $f(\mathbf{R}_{x,y})$ is the descriptor of the neighbourhood region $\mathbf{R} \subset \mathbf{I}$ centered at (x, y) , and $d(\cdot, \cdot)$ is a measure of similarity to compare descriptors. In most cases, \mathbf{R} has the same size as \mathbf{P} . In this category, attention is paid to the description $f(\cdot)$ of the information rather than in the training data or the classification scheme used. The description of a region determines in great measure the robustness of a recognition process facing viewpoint and illumination changes. Consequently, most researchers concentrate their efforts on obtaining invariant descriptors. Mikolajczyk and Schmid [19] classify the descriptors among the following categories: templates [1–6, 11], distributions [7, 8, 10, 20–29], Fourier [30] and Gabor transform, image derivatives [31], oriented filters [32, 33] and generalised moment invariants [34].

Among these descriptors and the recognition strategies used in those works, some have been chosen, because of their special relation to the method proposed here, and will be explained more in depth. The remainder of this section is structured as follows. Firstly, template and distribution descriptors are explained. Secondly, various strategies for region recognition within an image are described.

2.1 Template and distribution descriptors

Two descriptors have been used extensively for recognition purposes and, more specifically, in tracking applications [9]. These descriptors are based on templates and distributions. *Templates* are ordered arrays of the pixel values of a region, whereas *distribution* descriptors are arrays containing a discrete distribution of the information of a region.

Templates have two main advantages. First, the simplicity of construction of this descriptor. Second, the spatial information of the region is kept. The counterpart of this advantage is the high sensitivity to viewpoint and illumination changes. Several improvements of this simple matching technique exist in literature [2, 4, 6, 13]. Lewis [2] describes a fast computation of the cross-correlation with a normalisation invariant to illumination changes. However, this technique still lacks viewpoint robustness. [4, 6, 13] explore the parameterisation of the geometrical transformation that a patch suffers and, in this way, extend the robustness to rotation, translation and also scale changes.

A widely used distribution descriptor is the *histogram*. A histogram is an array that models the true distribution by counting the occurrences of pixel values that fall into each bin (which encompasses a range of values). Different information can be used for histogram-based descriptors, e.g. gray-scale [10], colour [7, 8, 22, 25], and gradient [26, 35]. Another distribution descriptor is the *signature* presented in [23]. Instead of using fix-sized bins, variable size is used for a better representation of the distribution space. Moreover, the length of the descriptor itself is also variable and depends on the complexity of the described image. In this way, more accurate (and larger) descriptors are obtained for complex images.

Histograms have opposite advantages and drawbacks when compared to templates. More concretely, histograms lose spatial information while viewpoint invariance can be achieved by construction. Several attempts at combining spatial and distribution information exist, e.g., [8, 10, 20, 21, 24, 26, 27, 29]. Comaniciu *et al.* [8] use a convex monotonic decreasing kernel that weights the contribution of pixels to the histogram. One advantage of this kernel is that the influence of peripheral pixels is lessened. Peripheral pixels are the least reliable, being often affected by occlusion and background (for instance, in tracking environments) and viewpoint changes (for instance, rotations). Lowe [26] uses the spatial distribution of gradient histograms in what is called Scale Invariant Feature Transform (SIFT). One of the achievements of SIFT is the high viewpoint invariance, gained partially by normalising histograms with respect to the dominant direction of the patch. Georgescu and Meer [27] use four oriented kernels that, in addition to spatial information, provide orientation data at 0, 45, 90 and 135 degrees. Adam *et al.* [10] employ multiple patches, each with a single histogram, to represent an object. The spatial arrangement of the patches overcomes the loss of spatial knowledge. Other examples of integration of spatial information into statistical descriptors are the co-occurrence matrices [20], colour correlograms [21] and multi-resolution histograms [24], which are especially used as global features for image indexing and retrieval, and, more recently, intensity-domain spin images [28] and spatiograms [29]. The region descriptor proposed in this work also combines spatial and distribution information as described in Section 3.2.

2.2 Matching strategies

The strategy to locate and match regions inside an image varies depending on the application and often also on the complexity of the descriptor or the measure of similarity used. However, each strategy is not solely related to one descriptor. In other words, for each strategy different descriptors can be used. Three main strategies can be identified in literature, namely, point correspondence, line-search and window-search matching.

In applications such as point correspondence [3, 26, 27, 31], the connection of points in two or more images, representing the same real point, is targeted. Consequently, only locations with high repeatability are considered. Since these regions or patches feature specific properties, they are called *interest*

(or feature) points or regions. For a review of interest region detectors the reader is referred to [36]. Once the detection of possible candidates (usually a large amount of points) in each image is done, a pair-wise match has to be set. Therefore, the similarity is only computed between pairs of interest points. For a large amount of points, this process is usually computationally complex. Nevertheless, methods to efficiently obtain the correct matches exist.

In some cases, there is no knowledge of specific locations (points in the image) that may match the region that is sought. However, in such situations, it is usually possible to determine an area of the test image which might contain a match. Two closely related strategies exist to deal with such situations, namely, line and window search.

In *line-search* matching, the goal is to maximise the similarity $d(\cdot, \cdot)$ between the patch and different points of an image. This process is done in several iterations starting at a known position. At each iteration the similarity between the patch and the neighbourhood of a given point is computed. Given this result it is possible to find another close location where the expected similarity is higher. The process is ended when the similarity is enough for the purposes of the application. Examples of this strategy can be found in [4, 6, 8, 29].

In *window-search* matching, the similarity is computed at each point in a test image. The result of this procedure is a *similarity map*. The computational power needed to build a similarity map is proportional to the size of the map. Due to this fact, it is often applied only when the descriptor at each point is computed rapidly or the size of the map is relatively small [1, 2, 5, 10, 11, 22, 30]. The RCM proposed here also produces a similarity map and hence is close to this sort of matching strategy.

Specific examples of fast rotation invariant matching with an exhaustive search are [37, 38]. Fredriksson *et al.* [37] use an orientation invariant descriptor (colour histogram), to locate points with high probability of match. Although this method is faster than commonly used cross correlation by FFT, histograms are not efficiently computed in this work. Ullah *et al.* [38] use the gradient orientation of each point of an image patch (forming what is called orientation codes) in a two step strategy. First, orientation code histograms (OH) are used to estimate the orientation of a patch in each point of an image. Second, orientation code matching (OCM) at the right orientation is applied only to the best histogram matches. This independent work differentiates from the method proposed here in one main contribution. This is, the OC is built only upon the extracted patch at a single orientation achieving less invariance to rotations than our descriptor. This is translated in a poorer performance as seen in Section 4.4.

For the sake of completeness, a brief note on similarity measures is also given hereafter. As said before, the matching performance is also related to the similarity $d(\cdot, \cdot)$ used to compare descriptors (see Equation (1)). For instance, cross correlation [2, 3, 11] and sum of squared distances (SSD) [1, 4, 6] are commonly used for template matching. In the case of histogram matching, more possibilities exist. The Bhattacharyya distance [39] and the Earth Mover's Distance (EMD) [23] have attracted the attention of researchers because of their discriminative and illumination invariance properties, respectively. Examples of the former applied to tracking frameworks can be found in [7, 8], whereas the latter has been applied both to indexing [23] and tracking [10]. For a more extended description of distribution similarity metrics, the reader is referred to [40].

3 Proposed method

As said in the previous section, most descriptors are focused on invariance losing interesting available information about the viewpoint. However, not being invariant but discriminative can be an additional asset. In other words, instead of building descriptors that do not vary with different conditions, one may consider building descriptors whose variation with different conditions is known. The technique

proposed here focuses on this idea. More precisely, the direction information obtained from the gradient is kept. This information is used to provide invariance to rotation in the matching process, together with an estimate of the actual rotation between the patch and the image. This dual process leads to the Rotation Correlation Map (RCM).

In this section, an overview of the method, the descriptor used for the recognition of a patch and the RCM are described.

3.1 Overview

Prior to describing the actual method to obtain the RCM and the descriptor tailored for it, a preliminary conceptual overview is given here. The goal of the RCM is to know the correlation and orientation of a patch with an image, at each point of this image. A straight approach to solve this problem is as follows. First, N versions of the patch rotated at different angles ($k \cdot 360^\circ/N$ with $k = 0, \dots, N - 1$) are generated. Then, a window-search template matching of each version with the image is computed. At each point, one obtains N levels of correlation. The maximum indicates which version is the most correlated, and, consequently, the approximate orientation of the original patch with respect to the image. Matching N templates has however a high computational cost (growing with N and the size of the template) so an efficient approach to achieve this goal is necessary.

The RCM is an efficient method to solve this problem. Instead of computing the correlation for each version, we estimate first which version has the highest probability of being the adequate to maximise the level of correlation. To select this version, the rotation of the patch with respect to the image has to be determined. This rotation is determined by comparing the direction of the patch and the direction of the region \mathbf{R} . The direction of a point is obtained from the orientation gradient at that point. Therefore, the direction of a region can be obtained with the orientation gradient histogram of that region. The comparison of histograms could be done at each point in the image. However, we opt for a more efficient pre-search based on gradient magnitude. The histogram similarity is computed only at those points with alike magnitude. Then, a fast and efficient strategy to perform such histogram matching consists in taking advantage of the Integral Histogram [41]. The result of histogram matching is twofold. On the one hand, points with high histogram similarity are found. On the other, the rotation is estimated at each point. With this information, it is possible to compute the correlation only at those points that have high probability of match and, in addition, compute this correlation with the closest rotated version, considerably reducing the total computation cost. This idea is developed and detailed in the remaining of this section.

3.2 Region description

Section 2 gives an idea of the vast number of different region descriptors available in literature. As stated before, two main problems have to be addressed in recognition, namely, illumination and viewpoint changes. These two issues are tackled simultaneously by using the gradient information. On the one hand, the gradient has little sensitivity to illumination changes. On the other, we propose a descriptor that addresses the viewpoint problem concentrating on rotation robustness and, at the same time, provides orientation information of the region it describes. This particular information is used to identify the rotation that a patch has undergone when detected in another image. This is the origin of the *discriminative* characteristic of the descriptor.

Let us first analyse the behaviour of the gradient. From a theoretical point of view, the gradient has a continuous response to a continuous and derivable function. Suppose that a gradient orientation histogram of N bins is computed from the intensity information of an image patch \mathbf{P} . In this case, a rotation of the patch by δ degrees changes the values in the histogram. In particular, when $\delta = n \cdot 360/N$ where $n \in \mathbb{Z}$, the histogram would be exactly equal to a perfect shift, and the shift in bins would be equal to n . However, this ideal case is not fulfilled in reality.

Following the observation that histograms change with different orientations, we propose to generate rotated versions of a patch and, from these versions, create a single histogram that can deal with rotations. As mentioned before, orientation histograms repeat approximately their shape every $\Delta = 360/N$ degrees. This can be exploited by aligning the histograms of versions rotated exactly by $k\Delta$ with $k \in \mathbb{Z}$.

The histogram descriptor is obtained as explained next. Firstly, N rotated versions of the patch \mathbf{P} to be matched are pre-computed with an angle of rotation of $n\Delta$ degrees (for $n = 0, \dots, N - 1$) where N is the number of bins. These versions are cropped so as to eliminate additional pixels introduced by the rotation, leading to a vector of rotated versions of the patch $\vec{\mathbf{P}}_i$, where i indexes the vector. Secondly, the gradient of each of these versions is computed at each point (x, y) as follows

$$\begin{aligned} dy(x, y) &= \vec{\mathbf{P}}_i(x, y + 1) - \vec{\mathbf{P}}_i(x, y - 1) \\ dx(x, y) &= \vec{\mathbf{P}}_i(x + 1, y) - \vec{\mathbf{P}}_i(x - 1, y) \\ \nabla_m(x, y) &= \sqrt{dy(x, y)^2 + dx(x, y)^2} \\ \nabla_\theta(x, y) &= \arctan(dy(x, y), dx(x, y)), \end{aligned} \quad (2)$$

where $\arctan(a, b)$ is a function that returns the inverse tangent of a/b in a range $[0, 2\pi[$, ∇_m is the magnitude and ∇_θ is the orientation of the gradient. Then, ∇_θ is quantised in N bins. In order to compact the statistical description of the patch and to reduce the effect of noise, the contribution of each point in $\nabla_\theta(x, y)$ to the corresponding bin is weighted by its magnitude $\nabla_m(x, y)$ (similar to the approach of Lowe [26]). Assuming b is a function that assigns its argument to the quantised space of bins, and δ is the Kronecker delta function, one obtains the expression for the histogram of a single rotated version $\mathbf{h}_{\vec{\mathbf{P}}_i}$

$$\begin{aligned} b &: [0, 2\pi[\mapsto \{0, \dots, N - 1\} \\ h_{\vec{\mathbf{P}}_i}(n) &= \sum_{\vec{\mathbf{P}}_i} \nabla_m(x, y) \cdot \delta[b(\nabla_\theta(x, y)) - n] \quad n = 0, \dots, N - 1 \end{aligned} \quad (3)$$

$$\mathbf{h}_{\vec{\mathbf{P}}_i} = [h(1), \dots, h(N)]_{\vec{\mathbf{P}}_i} \quad (4)$$

It is desirable that the weight of the peripheral pixels is lessened. However, applying a kernel (as presented by Comaniciu *et al.* [8]) is not possible with the integral histogram approach. The effect of the kernel is approximated by giving double weight to the central part of the patch. Redefining Equation (3),

$$\begin{aligned} h_{\vec{\mathbf{P}}_i}(n) &= \sum_{\vec{\mathbf{P}}_i} \nabla_m(x, y) \cdot \delta[b(\nabla_\theta(x, y)) - n] \\ &+ \sum_{\vec{\mathbf{P}}'_i} \nabla_m(x, y) \cdot \delta[b(\nabla_\theta(x, y)) - n] \quad n = 0, \dots, N - 1, \end{aligned} \quad (5)$$

where \mathbf{P}' is the central part of a patch \mathbf{P} . Finally, the global histogram of the patch is the mean obtained with the N histograms aligned according to their rotation.

$$\hat{\mathbf{h}}_{\vec{\mathbf{P}}_i} = [h(N - i), \dots, h(N - 1), h(0), \dots, h(N - 1 - i)]_{\vec{\mathbf{P}}_i} \quad (6)$$

$$\tilde{\mathbf{h}}_{\mathbf{P}} = \frac{1}{N} \sum_i \hat{\mathbf{h}}_{\vec{\mathbf{P}}_i} \quad (7)$$

Figure 1 shows an example for 16 bins with the original patch and its rotated versions with the corresponding histogram aligned accordingly.

This average of rotated versions gives a robust 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, experiments shows that, with enough bins, this descriptor is reliable even around $n\Delta + \Delta/2$ degrees (see Section 4).

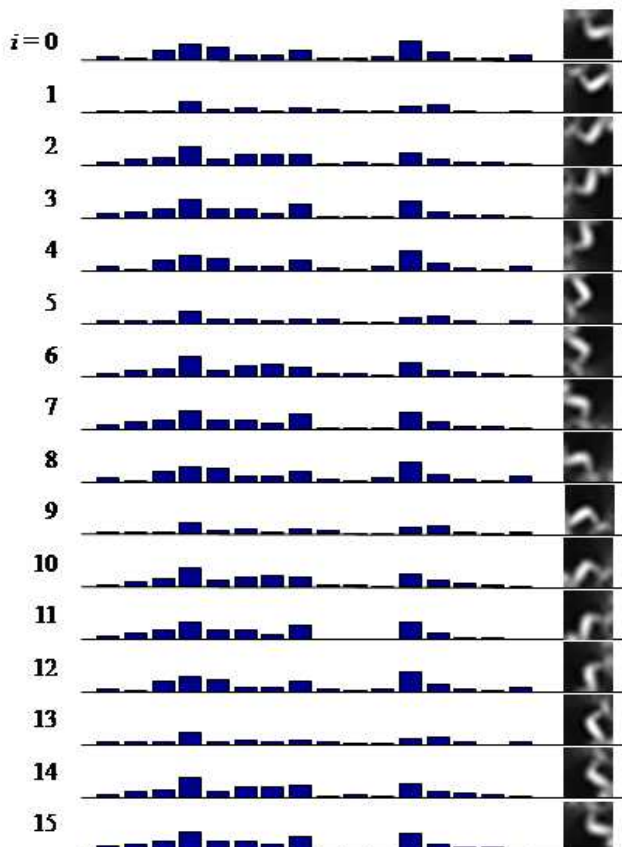


Figure 1: Example of histogram alignment with $N = 16$ bins. Central column: histograms aligned according to their rotation; right column: corresponding original patch and rotated versions.

The final region descriptor is composed of the global histogram $\tilde{\mathbf{h}}$, its variance σ^2 , its norm and the rotated versions of the template.

$$f(\mathbf{P}) = [\tilde{\mathbf{h}}_{\mathbf{P}}, \sigma_{\mathbf{P}}^2, \|\tilde{\mathbf{h}}_{\mathbf{P}}\|, \vec{\mathbf{P}}_0, \dots, \vec{\mathbf{P}}_{N-1}] \quad (8)$$

All of these elements are used in the matching procedure of the RCM explained below.

3.3 Rotation Correlation Map

The goal of the RCM is to obtain rotation invariance together with knowledge of the orientation of the patch that is sought with respect to the tested image. The question is, how are these two goals achieved by the RCM? In fact, the RCM is a vector-valued map of two components, namely, the rotation Θ , and the correlation Ψ

$$\mathbf{RCM}(x, y) = [\Theta(x, y) \ \Psi(x, y)]^T, \quad (9)$$

where T indicates vector transpose. In order to obtain these two maps, the process is divided in three hierarchical selection steps, each of them sorting out a set of most probable candidates. Firstly, an exhaustive gradient magnitude comparison is performed. Secondly, the candidates with highest magnitude similarity are kept for orientation gradient histogram matching. The similarity measure employed for histogram matching also provides an estimate of the rotation between the patch and the image, which results in Θ . Finally, the most similar histograms together with the rotation estimated at those positions are used in the template matching process that leads to Ψ . These steps are detailed in the remainder of this section.

3.3.1 Gradient magnitude matching

The norm of the histogram $\|\tilde{\mathbf{h}}_{\mathbf{P}}\|$ can be used as a simple feature to rapidly scan the image for similar candidates. From the construction of the histogram it can be found that,

$$\|\tilde{\mathbf{h}}_{\mathbf{P}}\| \simeq \sum_{\mathbf{P}} \nabla_m + \sum_{\mathbf{P}'} \nabla_m, \quad (10)$$

where \mathbf{P}' is the central part of the patch. Following this observation, we propose to compare this norm with each neighbourhood in a window-search strategy. This can be efficiently performed with the integral image [16] of the magnitude gradient ∇_m . Given a neighbourhood \mathbf{R} of a point in image I , the measure used to compare the norm is

$$d_m = \exp -\alpha \cdot \left(1 - \frac{\sum_{\mathbf{R}} \nabla_m + \sum_{\mathbf{R}'} \nabla_m}{\|\tilde{\mathbf{h}}_{\mathbf{P}}\|} \right)^2, \quad (11)$$

where α is a factor that weights this similarity according to the variance of the histogram. This factor is fixed, upon experimentation, to $\alpha = N/(1000 \cdot \|\sigma_{\mathbf{P}}^2\|)$. The points in the image that have a similarity $d_m > 0.9$ are kept as candidates for further matching. This set is called S_m .

In the worst case where similar magnitude is found all over the image, the number of candidates remains the same after this step. However, based on experiments, this simple selection criteria permits a reduction of the number of candidates by an average factor of 20.

3.3.2 Histogram matching

The gradient orientation histogram matching is applied to the candidates with similar histogram norm (S_m). Histograms are efficiently computed with the integral histogram approach [41]. With this method, computing the contribution of a single bin to the histogram can be performed with only four memory accesses. The gradient orientation histogram of a region in the image is obtained from the contribution of the quantised $\nabla_{\theta}(x, y)$ weighted with $\nabla_m(x, y)$ (as for the descriptor of the patch).

The similarity between the histogram of the patch $\tilde{\mathbf{h}}_{\mathbf{P}}$ and that of each candidate is computed with a custom measure to compare orientation histograms. Actually, this measure can be used with any sort of circular vector. We call it *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. Mathematically expressed

$$\text{CNED}(\mathbf{a}, \sigma_{\mathbf{a}}^2, \mathbf{b}) = [\hat{s}(\mathbf{a}, \sigma_{\mathbf{a}}^2, \mathbf{b}) \ d_{\hat{s}}(\mathbf{a}, \sigma_{\mathbf{a}}^2, \mathbf{b})]^T \quad (12)$$

$$\hat{s}(\mathbf{a}, \sigma_{\mathbf{a}}^2, \mathbf{b}) = \arg \min_s d_s(\mathbf{a}, \sigma_{\mathbf{a}}^2, \mathbf{b}) \quad (13)$$

$$d_s(\mathbf{a}, \sigma_{\mathbf{a}}^2, \mathbf{b}) = \sqrt{\sum_{i=0}^{N-1} \frac{(\mathbf{a}(i) - \mathbf{b}((i+s) \bmod N))^2}{\sigma_{\mathbf{a}}^2(i)}}, \quad (14)$$

where \mathbf{a} and \mathbf{b} are vectors of length N , s is the shift that takes a discrete value between 0 and $N - 1$, mod is the modulus function, and $\sigma_{\mathbf{a}}^2$ is the variance associated to vector \mathbf{a} . The result of this matching is hence a similarity score $d_{\hat{s}}$ and an estimate of the orientation of the patch $\hat{s} \cdot \Delta$ for each candidate.

Using this metric, the histogram matching step leads to a gradient histogram-based similarity map (GHSM)

$$\text{GHSM}_{\mathbf{P}, \mathbf{I}}(x, y) = \begin{cases} d_{\hat{s}}(\tilde{\mathbf{h}}_{\mathbf{P}}, \sigma_{\mathbf{P}}^2, \mathbf{h}_{\mathbf{R}_{x,y}}) & (x, y) \in S_m \\ 0 & \text{otherwise,} \end{cases} \quad (15)$$

where $\mathbf{h}_{\mathbf{R}_{x,y}}$ is the histogram computed for a region centered at (x, y) (see Equations (4-5)). In addition, this computation obtains the map of estimates of the rotation between the patch and the image

$$\Theta_{\mathbf{P}, \mathbf{I}}(x, y) = \begin{cases} \hat{s}(\tilde{\mathbf{h}}_{\mathbf{P}}, \sigma_{\mathbf{P}}^2, \mathbf{h}_{\mathbf{R}_{x,y}}) & (x, y) \in S_m \\ \text{undefined} & \text{otherwise.} \end{cases} \quad (16)$$

3.3.3 Template matching

The magnitude and the orientation histogram discard many unrelated points but the result is still not selective enough (as seen below in Section 4). Spatial intensity information (template) is used as a further selection criterion.

Template matching is done using a NCC between the templates (\mathbf{R} , neighbourhood pixel intensity) centered at those points with high histogram similarity ($S_h \subseteq S_m$) and the corresponding template of the patch $\vec{\mathbf{P}}_{\hat{s}}$

$$NCC(\vec{\mathbf{P}}_{\hat{s}}, \mathbf{R}) = \frac{\sum \sum (\mathbf{R} - \bar{\mathbf{R}}) \cdot (\vec{\mathbf{P}}_{\hat{s}} - \bar{\vec{\mathbf{P}}}_{\hat{s}})}{\sqrt{\sum \sum (\mathbf{R} - \bar{\mathbf{R}})^2 \cdot \sum \sum (\vec{\mathbf{P}}_{\hat{s}} - \bar{\vec{\mathbf{P}}}_{\hat{s}})^2}}, \quad (17)$$

where \bar{R} is the average value of R . By subtracting this mean value, the result is invariant to uniform illumination changes. In order to perform this computation fast, the integral image and image square of the tested image are computed so that $\sum \sum (\mathbf{R} - \bar{\mathbf{R}})^2$ can be computed in only a few memory accesses [2]. Additionally, more efficiency is gained by computing $\sum \sum (\vec{\mathbf{P}}_{\hat{s}} - \bar{\vec{\mathbf{P}}}_{\hat{s}})$ prior to template matching.

A correlation map Ψ can be built using the result of matching the templates of the candidates in S_h . The map takes value 0 everywhere except at the location of these candidates, where the value $\in [0, 1]$ is the NCC computed as described before

$$\Psi_{\mathbf{P}, \mathbf{I}}(x, y) = \begin{cases} NCC(\vec{\mathbf{P}}_{\Theta(x,y)}, \mathbf{R}_{x,y}) & (x, y) \in S_h \\ 0 & \text{otherwise.} \end{cases} \quad (18)$$

3.4 Computational complexity

The processing time needed for an image determines the applicability of the technique to real-time environments. This section discusses the computational complexity of the proposed algorithm and possible ways to lessen it.

The description of a patch takes three steps: the creation of the rotated versions, creation of each individual histogram (one per version) and, finally, alignment and descriptor computation. The first step can be performed very rapidly using the processing power of a graphic card and the last two are proportional to the size of patches, which is often very small. Hence, one advantage of this type of descriptors is that they can be computed on-the-fly. Therefore, in a tracking application it would be possible to add new regions to track at run-time.

The RCM computation is logically where most of the process elapses. Each step, separated in consecutive order, gives rise to the following cost.

- The computation of the gradient information (magnitude and orientation) and the integral histogram is done only once and is proportional to the size of the image ($W \cdot H$ pixels).
- The exhaustive magnitude comparison is performed at each point in the image and hence is also proportional to $W \cdot H$.
- The N bins-histogram is calculated in only $4N$ memory accesses and additions (independently of the size of the patch). This is the great advantage of the integral histogram over conventional methods (see [41] for a complete complexity derivation). This process is performed on a limited number of candidates $|S_m|$ with similar histogram norms.
- Each histogram $\mathbf{h}_{\mathbf{R}_{x,y}}$ is compared to the histogram of the patch $\tilde{\mathbf{h}}_{\mathbf{P}}$ using the CNED in N^2 operations.

- The template matching is proportional to the size of the patch and the number of candidates $|S_h|$ kept after the histogram matching.

To summarise, the matching step takes roughly $W \cdot H + (4N + N^2) \cdot |S_m| + (5 + W_{\mathbf{R}} \cdot H_{\mathbf{R}}) \cdot |S_h|$ operations, where $W_{\mathbf{R}} \cdot H_{\mathbf{R}}$ is the size of the patch.

Although the number of candidates $|S_h|$ plays a role in the processing speed, our experimentation has shown that three parameters determine the rapidity of the algorithm, namely, N , W and H . As shown in Section 4, the number of bins determines the results of the system, whereas the performance is independent of W and H . In addition, $|S_h|$ is in accordance to these two parameters. Consequently, W and H should be decreased in case fast performance is needed.

In some applications, there is a rough knowledge of the area, inside an image frame, where the patch may lie. In these cases, W and H can be drastically reduced. This is the case for tracking applications where the intra-frame motion can be predicted and the search region is known from the previous location. For 2D tracking environments, a more detailed description with additional advantages introduced by the RCM is given in Section 5.

4 Experiments

This section describes the evaluation of the proposed method. Firstly, the test set used to assess the performance is presented. Secondly, the correlation accuracy of the RCM is tested on its own. Thirdly, the orientation accuracy is discussed. Finally, a comparison with other similar techniques, both in terms of performance and computational cost, concludes this section. The methodology of each experiment is presented in its respective subsection.

4.1 Test set

The set of images used for testing is shown in Figure 2. The first two images are custom whereas the last six images are taken from the Visual Geometry Group database [42]. It can be seen that this set has textured regions and, in many cases, similarity between these regions.

There are two key parameters in the method: the number of bins in the histogram N and the number of candidates chosen from the histogram matching step $|S_h|$. Experiments are run on 10, 16 and 20 bins to give an approximate idea of a lower and upper performance bounds. The number of bins determines the value of Δ (see Section 3.2) and, hence, the performance of the method. More concretely, the RCM is expected to work better for rotations around $k\Delta$ than around $k\Delta + \Delta/2$ (with $k = 0, \dots, N - 1$). In order to experiment with these best and worst scenarios, the images are rotated according to each histogram length. More concretely, the images are rotated 20 and 70 degrees for a histogram of 10 bins ($\Delta = 36^\circ$), and 10 and 70 degrees for 16 bins ($\Delta = 22.5^\circ$), and 20 bins ($\Delta = 18^\circ$). The method is also run on the original images (no transformation). The number of candidates in the set S_h extracted from the GHSM ranges from 1 to 500.

For each one of the original images, a set of patches is extracted. The method used for patch extraction is tailored for the matching method proposed. Kadir and Brady [43] indicate the convenience of using the same feature(s) to detect regions and describe them. In their work, they use the entropy of the histogram as a feature of saliency. Moreover, the size of the detected region is chosen at a peak of entropy. Here a similar strategy is adopted.

In the descriptor presented in Section 3.2, the most relevant feature is the gradient orientation histogram, as it determines in great measure the performance of the method. Actually, it is the richness of this feature that enables discrimination and orientation estimation. Consequently, regions with a rich gradient orientation histogram must be found in these images. It is desirable that the shape of a histogram has peaks and valleys and even more important, that those peaks and valleys have a



(a) Sunflowers, 300x225 pixels



(b) Cathedral, 300x200 pixels



(c) Bark, 320x214 pixels



(d) Bikes, 300x210 pixels



(e) Boat, 300x240 pixels



(f) Graf, 300x240 pixels



(g) Leuven, 300x200 pixels



(h) UBC, 300x240 pixels

Figure 2: Original images used for the experiments.

non-periodic shape. In this way, the CNED achieves the best performance. This "peakyness" η is measured directly with the values of the histogram h computed at a given region

$$\eta = \sum_i \|h(i) - h(i-1)\|. \quad (19)$$

The procedure followed to obtain the patches is performed as described next.

1. Harris corner points [44] are first selected. This detector finds points with a high probability of multi-peaked orientation histograms and, basically, with a high gradient magnitude.
2. For each point, regions are extracted at different sizes. More precisely, from a 10×10 pixels to a 20×20 pixels. This range of sizes is selected because it gives the best results.
3. The gradient orientation histogram is computed for each region. With the histogram, η is obtained.
4. The size that maximises η is kept and its patch extracted.
5. If the variance of the histogram is below $0.025/N$ (chosen upon experimentation) the candidate is ignored. This indirectly enforces an heterogeneous shape.

Once the patch is extracted, the descriptors can be built. After applying this extraction to the test set, 161 regions are obtained for histograms with 10 bins, 152 for 16 bins and 153 for 20 bins.

In the following evaluations, each patch is sought independently in the transformed images. Results are averaged for all the patches extracted.

4.2 Evaluation of the correlation accuracy

The RCM gives a level of correlation at each point in the image. It would be possible to simply show the correlation map Ψ and let the reader evaluate its accuracy. However, showing all the maps for each patch and test image is not practical. Hence, in order to determine the performance of the proposed map, a numerical quantity that summarises all the responses is needed. This value should clearly show when a good localisation of a patch inside an image is provided. We consider that a good localisation is fulfilled when a high correlation is achieved at the ground truth position of the patch in the test image. The level of correlation and the number of good localisations among all the patches determine the capacity of the algorithm at different conditions (different test images). Whether the algorithm is capable of locating a patch at the right place is not enough to characterise the accuracy. In addition, this metric should measure the discrimination between the right position and the remaining points. The level of correlation with points unrelated to the good location gives an idea of that distinction.

These ideas are summarised mathematically as follows. As seen before, the ground truth location of the patch inside a test image is necessary. In our evaluation, the transformation applied to the original images is known so this ground truth is available. Let us assume there is an operator $T(\mathbf{I})$ that performs a similarity transformation (rotation, translation, scaling) of an image \mathbf{I} . Now, suppose that a patch \mathbf{P} is extracted from \mathbf{I} and the centre of the patch is located at $(x_{\mathbf{P}}, y_{\mathbf{P}}) \in \mathbf{I}$. Given an image $\hat{\mathbf{I}} = T(\mathbf{I})$, and $(\hat{x}_{\mathbf{P}}, \hat{y}_{\mathbf{P}})$ being the correspondence of $(x_{\mathbf{P}}, y_{\mathbf{P}})$ into $\hat{\mathbf{I}}$, two values are computed: the maximum correlation near the ground truth

$$\psi_{\in \mathbf{G}} = \max \Psi_{\mathbf{P}, \hat{\mathbf{I}}}(x, y) \quad \forall (x, y) \in \mathbf{G}, \quad (20)$$

and the maximum correlation outside the neighbourhood of the ground truth

$$\psi_{\notin \mathbf{G}} = \max \Psi_{\mathbf{P}, \hat{\mathbf{I}}}(x, y) \quad \forall (x, y) \notin \mathbf{G}, \quad (21)$$

where \mathbf{G} is the region $\{\hat{\mathbf{I}}(x, y) | x \in [\hat{x}_{\mathbf{P}} - 1, \hat{x}_{\mathbf{P}} + 1] \text{ and } y \in [\hat{y}_{\mathbf{P}} - 1, \hat{y}_{\mathbf{P}} + 1]\}$. After image transformation and interpolation, it is possible that the original centre of the patch lies between two pixels. A 1 pixel

neighbourhood is set to account for this sub-pixel location. It must be pointed out that these two values do not cover all the information that could be visualised in the map. For instance, the location of the point in the map that gives $\psi_{\notin G}$ is discarded. As a matter of fact, this point could be just 2 pixels away from (\hat{x}_P, \hat{y}_P) or at a distant location. In most applications, the first case would not cause a problem whereas the second case would be much more relevant. However, by giving only a 1 pixel neighbourhood the accuracy of our method is clearly tested.

Figure 3 shows $\psi_{\in G}$ (solid line) and $\psi_{\notin G}$ (dashed line) averaged for all the patches in the test set. As it can be seen, the number of candidates kept from the histogram matching step $|S_h|$ has a direct

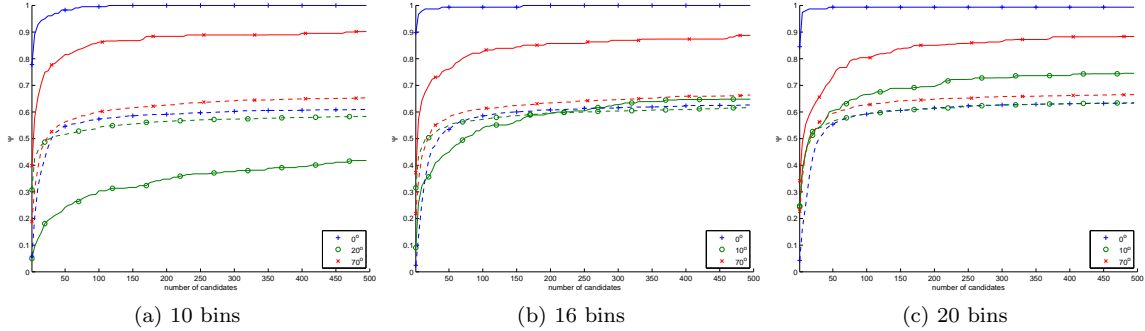


Figure 3: Mean $\psi_{\in G}$ (solid line) and $\psi_{\notin G}$ (dashed line). Histograms computed with 10, 16 and 20 bins.

influence on the performance. On the one hand, a small number of candidates (less than 100) is not reliable enough as in most cases $\psi_{\in G}$ is small. On the other hand, for a number of candidates greater than 150 in general, several observations arise.

- The performance is improved as the number of bins in the histogram grows. This is especially visible for rotations around $k\Delta + \Delta/2$. In this particular test set, these rotations are at 20° , 10° and again 10° , for 10, 16 and 20 bins, respectively (lines with circles in Figure 3).
- The performances achieved for rotations around $k\Delta$ (lines with '+' and 'x') are globally better than those achieved for $k\Delta + \Delta/2$, as expected (lines with circles).
- The correlation inside G is larger than the correlation outside ($\psi_{\in G} > \psi_{\notin G}$), achieving the desired discrimination.
- A high correlation is achieved near the ground truth using 20-bins histograms ($\psi_{\in G} \geq 0.7$ in the worst case, 10°).

It must be pointed out that the number of candidates $|S_h|$ is related to the size of the images. In this case, 150 candidates is around 0.25% of the mean area of each image.

4.3 Evaluation of the rotation accuracy

One of the enhancements brought by the proposed method is the estimation of the relative rotation of a patch when detected in a test image. Although directly related to Ψ , the accuracy of this estimation is analysed for a more complete assessment.

The estimated orientation is given by $\Theta(x, y)$ (see Eq. 16) multiplied by the factor Δ to obtain the degrees of rotation. In the matching process, $\Theta(x, y)$ is used to compute the correlation $\Psi(x, y)$ (see Eq. 18). The validity of $\Theta(x, y)$ could then be corrected with this resulting correlation. This process is not done in the RCM to keep complexity low. However, this reasoning is used here for the evaluation

of the orientation estimation. For each patch, the ground truth region $\mathbf{G}_{\hat{x}_{\mathbf{P}}, \hat{y}_{\mathbf{P}}}$ in \hat{I} where the patch \mathbf{P} lies is selected in both $\Theta(x, y)$ and $\Psi(x, y)$. Then the orientation of each patch is computed as a weighted sum

$$\theta_{\mathbf{P}, \hat{\mathbf{I}}} = \frac{\sum_{\mathbf{G}} \Psi(x, y) \cdot \Theta(x, y) \cdot \Delta}{\sum_{\mathbf{G}} \Psi(x, y)}. \quad (22)$$

Recalling from Section 3.3.2, the shift stored in Θ is circular by definition. This means that in the same way that a rotation is periodic every 360° , the shift is periodic every N ($\Delta = 360/N$). As the rotations of the test set are in the range $[-180^\circ, 180^\circ]$ it is more convenient to express the shift in the range $[-N/2, N/2]$. This fact is considered and the values of Θ inside the region G are changed accordingly. In order to compact the results, the orientation for all patches is grouped according to the orientation of the test image used. Results for all images in the test set are averaged together. Table 1 shows this mean estimation for a given number of bins and rotation angle of $\hat{\mathbf{I}}$ with respect to \mathbf{I} . The accuracy is logically limited by the number of bins. In other words, the estimation is around a

	Orientation [degrees]			
	0	10	20	70
10 bins ($\Delta = 36^\circ$)	0.64	n/a	28.08	70.54
16 bins ($\Delta = 22.5^\circ$)	0.65	8.35	n/a	66.95
20 bins ($\Delta = 18^\circ$)	0.28	12.61	n/a	70.13

Table 1: Orientation θ estimated for each combination of histogram length and orientation of the transformed image.

multiple of Δ . These results are coherent with previous experiments. The estimation around $k\Delta + \Delta/2$ is less accurate, which is especially visible when only 10 bins are used.

It is important to underline the extra achievement of this method. Histogram descriptors are built upon rotated versions of a patch (see Section 3.2). These rotations are performed with respect to the centre of the patch. However, the rotations performed on the test images are applied with respect to their own centre, producing a distortion on the patches that is different from an exact rotation from their respective centres. Therefore, our method demonstrates robustness in front of these small similarity transformations (not only a rotation) of each patch.

More results of the RCM with other similarity transformations (including small image scaling) have been reported in [35]. In this case, a wide range of rotations (approximately a range of 180 degrees) is covered. Both good accuracy in the location of the patch and in the orientation at each frame of a video sequence is achieved.

4.4 Comparison with other similar techniques

This section assesses the performance of the RCM in comparison to other similar techniques and is structured as follows. Firstly, the techniques compared are described. Secondly, the evaluation methodology is explained. Finally, results are depicted and discussed.

The matching techniques compared are listed below.

A rotation-exhaustive template matching (NCC-R) The NCC is computed at each point of the tested image and for each of the N rotated versions. Only the best result of the N correlations computed at each point is kept. Hence, it is invariant to rotation transformations with the limitation of the number of versions used.

A gray-level intensity histogram matching (IHM) The histogram of the intensity is compared at each point. In the intensity histogram descriptor, the central part of each patch has double weight to lessen the effect of peripheral pixels as for $\mathbf{h}_{\mathbf{P}}$ (see Equations (4-5)). Moreover, the

descriptor of the patch that is sought is built upon the mean of intensity histograms of the N rotated versions. Although the intensity histogram of a single version should already be highly invariant to rotations, this averaging eliminates possible variations among rotated versions and hence produces a fairer comparison to the method proposed in this work. The similarity measure used is the Euclidean distance. Results using the Bhattacharyya distance [39] showed similar behaviour and, consequently, are not shown here to avoid redundancy.

A gradient orientation histogram matching (GHM) The comparison of gradient orientation histograms is applied as in the RCM (see Section 3.3.2). The difference is that the GHSM (Equation (15)) is computed in this case at each point in the image and not only at the set of candidates S_m .

OC histogram matching followed by OC matching (OH+OCM) Recalling from Section 2.2, this technique consists in a two step strategy [38]. Firstly, orientation code histograms (OH) are used to estimate the orientation of a patch in each point of an image. Second, orientation code matching (OCM) at the right orientation is applied only to the best histogram matches. The reason for comparing this strategy is threefold:

- to see the results for a larger data set than the one used in [38];
- to evaluate the improvement introduced by the proposed rotation-discriminative descriptor, which is one of the main differences with the OH+OCM method;
- to analyse the efficiency of this hierarchical search approach when compared to the RCM.

The proposed Rotation Correlation Map (RCM) In particular, the correlation map Ψ .

Each one of these matching techniques is computed in a different manner targeting a broad range of possibilities. On the one hand, two sorts of information are used, either pixel intensity (NCC-R and IHM) or gradient (GHM), and in the case of OH+OCM and RCM, a combination of both. And on the other hand, histogram matching (GHM and IHM) is compared to template matching (NCC-R) and to the mixed approaches, namely, OH+OCM and RCM.

In the evaluation of Section 4.2, the correlation level is considered as a measure of performance. Another possibility is considered here. The purpose of all the compared methods is to find matches. A match is expected at a peak in the similarity. A correct match is found when the peak coincides with the ground truth. An incorrect match is found at a peak unrelated to the ground truth. These ideas are translated into the concepts of *true positives* and *false positives*, respectively. This nomenclature is often used in Receiver Operating Characteristic (ROC) curves [45]. The performance can be given by these two values as follows: the higher the number of true positives and lower the number of false positives, the better is the obtained performance.

The level of similarity that determines a match (or positive) is given by a range, i.e. maximum to minimum similarity. However, the techniques that are considered do not have the same range. Indeed, the values in the NCC-R and Ψ range from 0 to 1, whereas the IHM and GHM compute distances whose range is not known a priori. Nevertheless, it is possible to find a range that varies equivalently among the techniques compared. In order to find this equivalence, the values of each map are taken, independently, in descending order (highest to lowest similarity) regardless of the value itself. An equivalent level of similarity is found in this case by parsing each list of values. This procedure is found to give a fair comparison among the maps. Following this procedure of descending similarity, the true positives and false positives can be defined mathematically. Given an image \mathbf{I} , a patch \mathbf{P} extracted from \mathbf{I} , $\hat{\mathbf{I}} = T(\mathbf{I})$, and $(\hat{x}_{\mathbf{P}}, \hat{y}_{\mathbf{P}})$ being the correspondence of $(x_{\mathbf{P}}, y_{\mathbf{P}})$ into $\hat{\mathbf{I}}$, a *true positive* is

$$tp_{\mathbf{P}, \hat{\mathbf{I}}, t} = \begin{cases} 1 & \text{if } \exists (x, y) \in \mathbf{G} \mid d(f(\mathbf{P}), f(\mathbf{R}_{x,y})) > t \\ 0 & \text{otherwise,} \end{cases} \quad (23)$$

Conversely, a *false positive* is

$$fp_{\mathbf{P}, \hat{\mathbf{I}}, t}(x, y) = \begin{cases} 1 & \text{if } (x, y) \notin \mathbf{G} \mid d(f(\mathbf{P}), f(\mathbf{R}_{x,y})) > t \\ 0 & \text{otherwise,} \end{cases} \quad (24)$$

where t takes values in the range $[\max d(f(\mathbf{P}), f(\mathbf{R}_{x,y})), \min d(f(\mathbf{P}), f(\mathbf{R}_{x,y}))]$. Similarly to the evaluation in Section 4.2, these values do not give a perception of distance in pixels from the location of a false positive to the ground truth.

The comparison methodology is applied to the test set. The number of rotated versions used in the NCC-R method is the same as the number of bins in the histograms of the other maps. Consequently, its results may vary with different bin numbers.

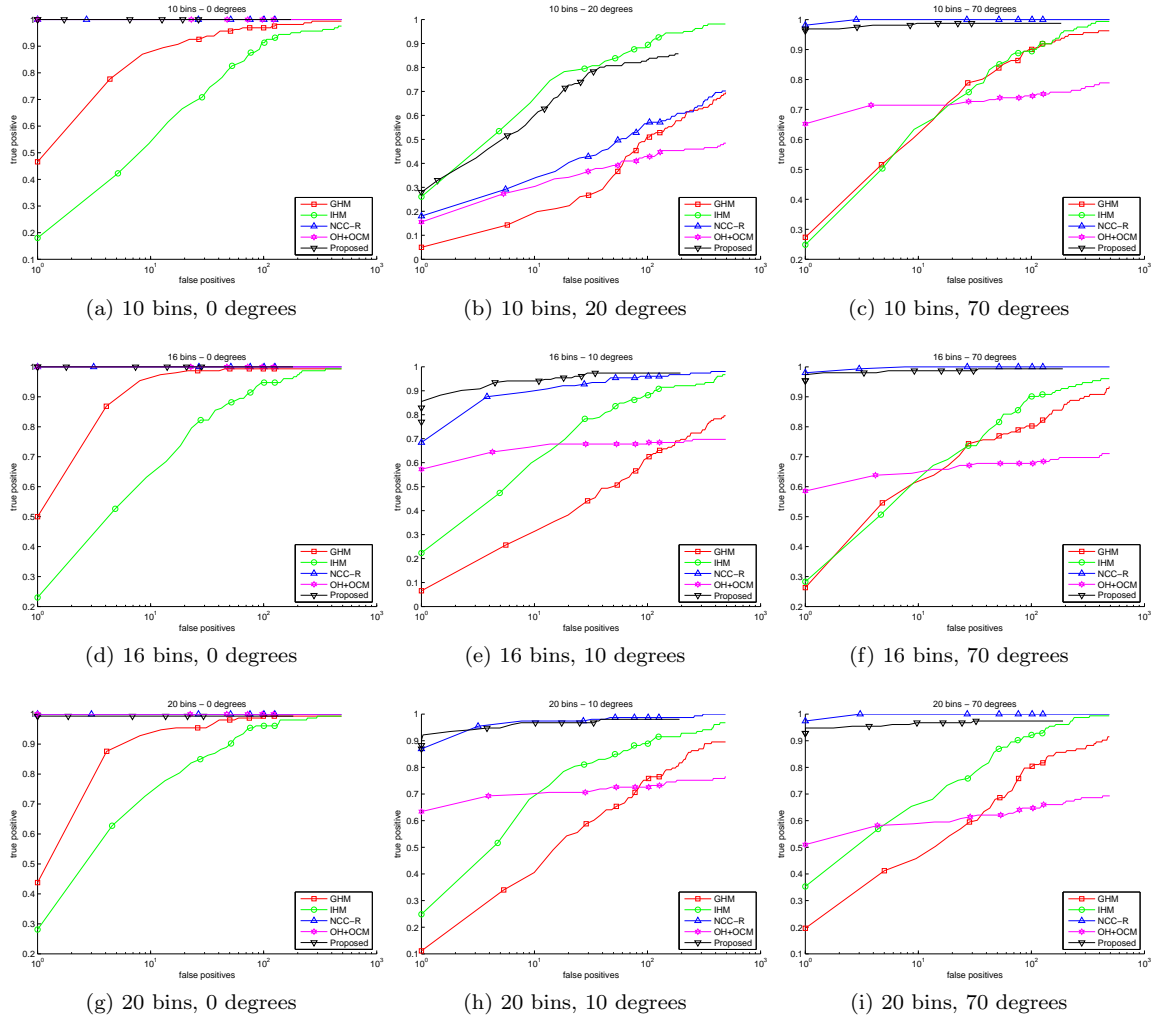


Figure 4: Average true positive (tp) and false positives (fp) among all the patches. Rotation angle: 0 degrees (left column), $\sim k\Delta + \Delta/2$ (central) and $\sim k\Delta$ (right).

The response of each similarity map is depicted in Figure 4. The NCC-R indicates a great performance almost independent of the number of candidates. This shows the high selectivity of this kind of map. In the case of the IHM, rotation invariance is evidenced by very similar results throughout the different cases. Moreover, for a small histogram length it achieves the best results when the rotation is around $k\Delta + \Delta/2$ (Figure 4b). A poorer selectivity is shown by the GHM as a large number of false positives is obtained in order to get a high probability of having a true positive. The results of the GHM are greatly improved when used as an input for further template matching as in RCM (especially visible as the number of bins grows). Furthermore, the proposed descriptor and similarity measure achieve the desired rotation discrimination and hence accurate matching. The OH+OCM [38] has lower performance probably due to its non-invariant nor robust descriptor. Using only a single version to build the histogram is not enough to effectively face the variations in the histogram due to rotations.

It could be argued that the selection of patches according to the matching method (see Section 4.1) biases the results favouring the RCM. In order to provide a baseline of performance, another set of image patches has been extracted. For each image of the original test set, 15 patches are extracted randomly at different points. The size of the patch is also selected randomly with the same range as before, namely, 10×10 to 20×20 pixels. The experiment is applied to this new set of patches obtaining the results depicted in Figure 5.

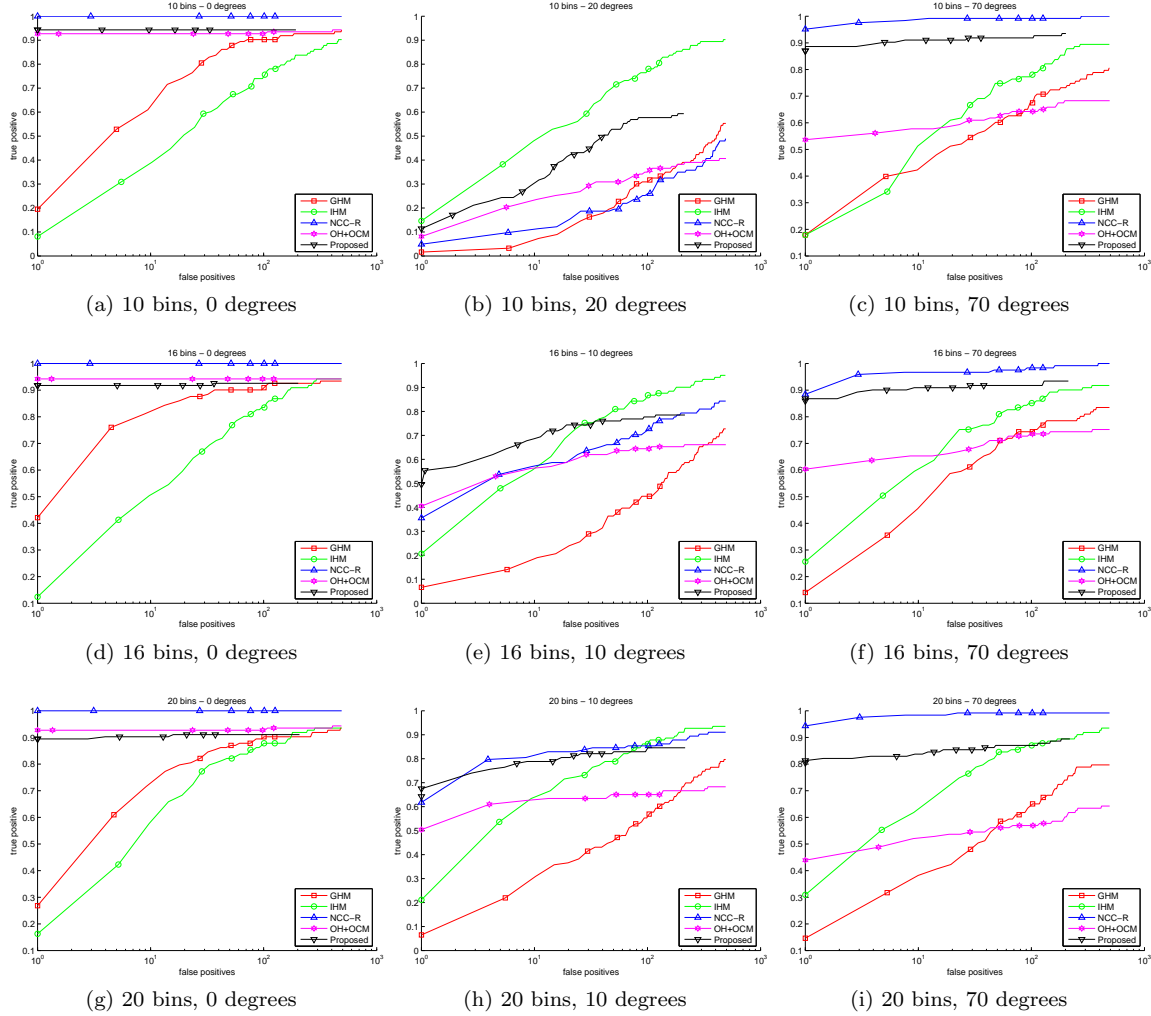


Figure 5: Average true positive (tp) and false positives (fp) among all the patches extracted randomly. Rotation angle: 0 degrees (left column), $\sim k\Delta + \Delta/2$ (central) and $\sim k\Delta$ (right).

Contrary to what could be expected, the new set of patches does not have a big impact and the results are very similar. One possible reason is that most images in the set are textured. However, it can be seen that patches without such texture (e.g., those extracted from the flat region of image Cathedral) induce a poorer performance almost in all the matching techniques compared.

5 Application to visual tracking

Registration of objects or mapping of environments is a necessary task for visual tracking. Visual tracking applications can be divided into bottom-up and top-down approaches [8]. Bottom-up approaches rely on the accurate representation of the tracked target, whereas top-down approaches are based on filtering techniques relying on a motion model of the target. Depending on the application

or scenario, attention is focused on either one or the other approach. For example, in cases where motion is relatively small, the descriptor used for object recognition is more relevant [5, 8, 10, 22]. Conversely, in cases where motion is larger, most trackers rely on multiple interest points with a simpler descriptor faster to match. This is the case of tracking techniques using interest points for filter update [3, 11, 12, 14]. In these works, template matching is often employed for point correspondence. When the camera rotates with respect to the normal of template, correlation fails and the update is incomplete. Molton *et al.* [13] presented a solution to overcome the problem of the correlation test for locally planar surfaces. By warping the template according to the current state of the filter, the warped version is closer to the appearance of the region in the current frame, thus leading to correct filter updates. Sim *et al.* [15] have attempted to combine both good representation and motion state filtering approaches. They use SIFT features [26] together with a particle filter [46] for camera tracking.

However, none of these techniques exploit directly the viewpoint determined at recognition level. We describe here a possible way to use the RCM to provide an extra input to a generic 2D filter-based tracking system. Assume that the filter of this generic tracker is governed by the following motion and measurement models, respectively,

$$\begin{aligned}\mathbf{x}_k &= u(\mathbf{x}_{k-1}, \mathbf{q}_{k-1}) \\ \mathbf{z}_k &= v(\mathbf{x}_k, \mathbf{r}_k),\end{aligned}\tag{25}$$

where $\mathbf{x} = (\hat{x}, \hat{y})$ is the state vector, \mathbf{z} the measurement vector, u and v are possibly non-linear functions, \mathbf{q} and \mathbf{r} are the process and measurement noises, respectively, and k is the current frame. The state vector represents the position of the target. The measurement \mathbf{z} used for filter update is the similarity map of the target with an area \mathbf{A}_k of the current frame containing the expected location of \mathbf{x}_k (often called the gating region). One straight possibility is to use the correlation part of the RCM. In this case, the likelihood of the filter is

$$p(\mathbf{z}_k | \mathbf{x}_k) \propto \Psi(x, y) \mid (x, y) \in \mathbf{A}_k.\tag{26}$$

Although the correlation map Ψ has shown good accuracy (see previous section), some false positives may still appear in the gating region which could induce erroneous updates. The question now is how to take advantage of the orientation part of the RCM to overcome this problem. The orientation of the neighbourhood of the target is determined by the tangential direction θ of the trajectory followed by the target $\mathbf{x}_{0:k} = (\hat{x}_{o:k}, \hat{y}_{o:k})$. Assuming that this direction can be obtained from the previous and current frames, one has

$$\theta_k = \arctan(\hat{y}_k - \hat{y}_{k-1}, \hat{x}_k - \hat{x}_{k-1}),\tag{27}$$

If the target's descriptor is initialised at a reference orientation θ_0 , then Θ of the current frame k is a valid measurement to correct the filter state

$$p(\mathbf{z}_k | \mathbf{x}_k) \propto \Psi(x, y) \cdot \exp\left(-[(\theta_k - \theta_0) - \Theta(x, y) \cdot 2\pi/N]^2\right) \mid (x, y) \in \mathbf{A}_k.\tag{28}$$

False positives can then be discarded in a straightforward manner using the state of the filter.

6 Conclusions

The Rotation Correlation Map (RCM) has been described. It is composed of a rotation map and a correlation map. The first map provides an estimate of the rotation of the patch that is sought, with respect to a tested image, at each point of the latter. The second map indicates the level of correlation of this patch, rotated at each point according to orientation map, with the image. In order to speed-up the computation of these two maps, the following process is done. First, a fast gradient orientation histogram matching is applied. The goal of this matching step is twofold. On the one hand, obtain an initial map of most probable locations of high correlation. On the other hand, obtain the rotation map. With these two maps, a template matching step is performed on those points selected as most probable. The result of this template matching is the correlation map.

The performance of the proposed method has been tested on several images for various different histogram lengths. In particular, the discrimination of the map and the accuracy of the rotation estimated are proven. In addition, the RCM outperforms other related similarity maps for patch recognition as shown in a comparative experiment.

An application to visual tracking has also been presented. The orientation map is used to refine the measurement-to-track association.

Future paths of research will focus on more generic affine transformations.

7 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. Finally, we are grateful to Yannick Maret for fruitful discussions and insights.

References

- [1] J. Shi and C. Tomasi, “Good features to track,” in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 1994, pp. 593–600.
- [2] J.R. Lewis, “Fast template matching,” *Vision Interface*, pp. 120–123, 1995.
- [3] I.J. Cox and S.L. Hingorani, “An efficient implementation of Reid’s multiple hypothesis tracking algorithm and its evaluation for the purpose of visual tracking,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 18, no. 2, pp. 138–150, Feb 1996.
- [4] G.D. Hager and P.N. Belhumeur, “Real-time tracking of image regions with changes in geometry and illumination,” in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 1996, pp. 403–410.
- [5] P. Fieguth and D. Terzopoulos, “Color-based tracking of heads and other mobile objects at video frame rates,” in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Washington, DC, USA, 1997, pp. 21–27.
- [6] F. Jurie and M. Dhome, “Hyperplane approximation for template matching,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 996–1000, 2002.
- [7] P. Perez, C. Hue, J. Vermaak, and M. Gangnet, “Color-based probabilistic tracking,” in *Proc. European Conference on Computer Vision (ECCV)*, London, UK, 2002, pp. 661–675, Springer-Verlag.
- [8] D. Comaniciu, V. Ramesh, and P. Meer, “Kernel-based object tracking,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 25, no. 5, pp. 564–577, 2003.
- [9] B. Deutsch, Ch. Graessl, F. Bajramovic, and J. Denzler, “A comparative evaluation of template and histogram based 2d tracking algorithms,” *Lecture Notes in Computer Science*, vol. 3663, pp. 269–276, 2005.
- [10] A. Adam, E. Rivlin, and I. Shimshoni, “Robust fragments-based tracking using the integral histogram,” in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2006, vol. 1, pp. 798–805.
- [11] Q. Zheng and R. Chellappa, “Automatic feature point extraction and tracking in image sequences for unknown camera motion,” in *Proc. Intl. Conf. on Computer Vision (ICCV)*, 1993, pp. 335–339.
- [12] A.J. Davison, “Real-time simultaneous localisation and mapping with a single camera,” in *Proc. Intl. Conf. on Computer Vision (ICCV)*, 2003.

- [13] N. Molton, A. Davison, and I. Reid, “Locally planar patch features for real-time structure from motion,” in *Proc. British Machine Vision Conference (BMVC)*, 2004.
- [14] M. Pupilli and A. Calway, “Real-time camera tracking using a particle filter,” in *Proc. British Machine Vision Conference (BMVC)*, September 2005, pp. 519–528.
- [15] R. Sim, P. Elinas, M. Griffin, and J. J. Little, “Vision-based SLAM using the Rao-Blackwellised particle filter,” in *Proc. IJCAI Workshop on Reasoning with Uncertainty in Robotics (RUR)*, Edinburgh, Scotland, 2005, pp. 9–16.
- [16] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2001, vol. 1, pp. 511–518.
- [17] V. Lepetit, J. Pilet, and P. Fua, “Point matching as a classification problem for fast and robust object pose estimation,” in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2004, vol. 2, pp. 244–250.
- [18] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.
- [19] K. Mikolajczyk and C. Schmid, “A performance evaluation of local descriptors,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005.
- [20] R.M. Haralick, K. Shanmugam, and I. Dinstein, “Textural features for image classification,” *IEEE Trans. on Systems, Man., and Cybernetics*, vol. 3, no. 6, pp. 610–621, 1973.
- [21] J. Huang, S. R. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih, “Image indexing using color correlograms,” in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Washington, DC, USA, 1997, pp. 762–768.
- [22] S. Birchfield, “Elliptical head tracking using intensity gradients and color histograms,” in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 1998, pp. 232–237.
- [23] Y. Rubner, C. Tomasi, and L. J. Guibas, “The earth mover’s distance as a metric for image retrieval,” *Intl. Journal of Computer Vision*, vol. 40, no. 2, pp. 99–121, 2000.
- [24] E. Hadjidemetriou, M.D. Grossberg, and S.K. Nayar, “Spatial information in multiresolution histograms,” in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2001, vol. 1, pp. 702–709.
- [25] T.-L. Liu and H.-T. Chen, “Real-time tracking using trust-region methods,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 26, no. 3, pp. 397–402, 2004.
- [26] D. Lowe, “Distinctive image features from scale-invariant keypoints,” *Intl. Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
- [27] B. Georgescu and P. Meer, “Point matching under large image deformations and illumination changes,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 26, no. 6, pp. 674–688, 2004.
- [28] S. Lazebnik, C. Schmid, and J. Ponce, “A sparse texture representation using local affine regions,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1265–1278, 2005.
- [29] S. Birchfield and S. Rangarajan, “Spatiograms versus histograms for region-based tracking,” in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, San Diego, California, June 2005, vol. 2, pp. 1158–1163.
- [30] A. J. Fitch, A. Kadyrov, W. J. Christmas, and J. Kittler, “Fast robust correlation,” *IEEE Trans. on Image Processing*, vol. 14, no. 8, pp. 1063–1073, 2005.
- [31] C. Schmid and R. Mohr, “Local grayvalue invariants for image retrieval,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 19, no. 5, pp. 530–535, May 1997.

- [32] W. Freeman and E. Adelson, “The design and use of steerable filters,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 13, no. 9, pp. 891–906, 1991.
- [33] G. Carneiro and Allan D. Jepson, “Phase-based local features,” in *Proc. European Conference on Computer Vision (ECCV)*, London, UK, 2002, pp. 282–296, Springer-Verlag.
- [34] L. J. Van Gool, T. Moons, and D. Ungureanu, “Affine/ photometric invariants for planar intensity patterns,” in *Proc. European Conference on Computer Vision (ECCV)*, London, UK, 1996, pp. 642–651, Springer-Verlag.
- [35] D. Marimon and T. Ebrahimi, “Orientation histogram-based matching for region tracking,” in *Intl. Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, 2007.
- [36] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool, “A comparison of affine region detectors,” *Intl. Journal of Computer Vision*, vol. 65, no. 1/2, pp. 43–72, 2005.
- [37] K. Fredriksson and E. Ukkonen, “Faster template matching without FFT,” in *Proc. IEEE Intl. Conf. on Image Processing (ICIP)*, 2001, vol. 1, pp. 678–681.
- [38] F. Ullah and S. Kaneko, “Using orientation codes for rotation-invariant template matching,” *Pattern Recognition*, vol. 37, no. 2, pp. 201–209, February 2004.
- [39] T. Kailath, “The Divergence and Bhattacharyya distance measures in signal selection,” *IEEE Trans. on Communications*, vol. 15, no. 1, pp. 52–60, 1967.
- [40] Y. Rubner, J. Puzicha, C. Tomasi, and J.M. Buhmann, “Empirical evaluation of dissimilarity measures for color and texture,” *Computer Vision and Image Understanding*, vol. 84, no. 1, pp. 25–43, 2001.
- [41] F. Porikli, “Integral histogram: a fast way to extract histograms in cartesian space,” in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2005, vol. 1, pp. 829–836.
- [42] Visual Geometry Group, Robotics Research Group, Department of Engineering Science, University of Oxford, ,” <http://www.robots.ox.ac.uk/~vgg/data.html>.
- [43] T. Kadir and M. Brady, “Saliency, scale and image description,” *Intl. Journal of Computer Vision*, vol. 45, no. 2, pp. 83–105, 2001.
- [44] C. Harris and M. Stephens, “A combined corner and edge detector,” in *Alvey Vision Conf.*, 1988, pp. 147–151.
- [45] T. Fawcett, “An introduction to ROC analysis,” *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.
- [46] M.S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, “A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking,” *IEEE Trans. on Signal Processing*, vol. 50, no. 2, pp. 174–188, Feb. 2002.