High-dimensional statistical measure for region-of-interest tracking

Sylvain Boltz, Eric Debreuve, Michel Barlaud

High-dimensional statistical measure for region-of-interest tracking

S. Boltz* É. Debreuve* M. Barlaud*

PREPRINT – Published in *IEEE Transactions on Image Processing*

Abstract

This paper deals with region-of-interest (ROI) tracking in video sequences. The goal is to determine in successive frames the region which best matches, in terms of a similarity measure, a ROI defined in a reference frame. Some tracking methods define similarity measures which efficiently combine several visual features into a probability density function (PDF) representation, thus building a discriminative model of the ROI. This approach implies dealing with PDFs with domains of definition of high dimension. To overcome this obstacle, a standard solution is to assume independence between the different features in order to bring out low-dimension marginal laws and/or to make some parametric assumptions on the PDFs at the cost of generality. We discard these assumptions by proposing to compute the Kullback-Leibler divergence between high-dimensional PDFs using the k-th nearest neighbor framework. In consequence, the divergence is expressed directly from the samples, *i.e.* without explicit estimation of the underlying PDFs. As an application, we defined 5, 7, and 13-dimensional feature vectors containing color information (including pixel-based, gradient-based and patch-based) and spatial layout. The proposed procedure performs tracking allowing for translation and scaling of the ROI. Experiments show its efficiency on a movie excerpt and standard test sequences selected for the specific conditions they exhibit: partial occlusions, variations of luminance, noise, and complex motion.

Keywords: Region-of-interest tracking, high-dimensional probability density function, nonparametric estimation, k-th nearest neighbor, Kullback-Leibler divergence.

1 Introduction

Tracking a region of interest (ROI) in a video is still a challenging task. Various high-level applications rely on tracking, *e.g.*, motion picture indexing, object recognition, video surveillance, audiovisual post-production... The problem can be defined as follows: a ROI is defined in a reference frame and the purpose is to determine in each subsequent frame the region which best matches the ROI in terms of a given similarity measure. Geometrically speaking, the two regions can be deduced from one another by an apparent motion usually restricted to a given model. Two classical similarity measures are the Sum of Squared Differences (SSD) and the Sum of Absolute Differences (SAD) between the reference ROI and a candidate region in a target frame.

1.1 A statistical approach

Similarity measures such as SSD and SAD impose a strict geometric constraint since the underlying residual is computed with a deterministic pixel-to-pixel correspondence between the reference ROI and the target region. In general, this apparent motion follows a rather simple model so that the estimation of its parameters remains well-posed. Therefore, it is not adapted to complex motions. Moreover, this type of similarity measures corresponds to implicit parametric assumptions on the residual probability density function (PDF) (resp., Laplacian and Gaussian for the two examples above [7]).

An alternative is to adopt a statistical point of view by building a PDF from the ROI and using it as a reference to be compared to a target PDF built from a candidate region by means of a similarity measure. Such statistical methods account for randomness and uncertainty in the observations, and therefore for complex motions. The PDFs can be defined in a radiometric space [16, 38], either in grayscale or color. However, to improve tracking accuracy, later developments tend to show that more information is required than just color. Different features were then integrated into the ROI PDF model, *e.g.*, recurring to filters such as spatial derivative filters [35, 10, 9], Gabor or wavelet filters [37], and temporal filters [10, 13]. A review of methods based on this framework proposed for segmentation was recently carried out [19].

While this increase of description features improves accuracy, their combination leads to high-dimensional PDFs. There exist efficient [40, 29] and fast [45] methods to estimate multivariate PDFs using Parzen win-

*Laboratoire I3S, UMR CNRS 6070, Sophia Antipolis, France (boltz,debreuve,barlaud@i3s.unice.fr).

dowing. However, due to the fixed cardinality of the data set, a limitation known as the curse of dimensionality [40] appears: as the dimension of the domain of definition of the PDFs gets higher, the domain sampling gets sparser. One can think of dilating the Parzen window [13] so as to ensure that it will enclose enough samples. However, the resulting PDF is oversmoothed. Another standard solution is to assume independence between the different features in order to bring out low-dimension marginal laws [10] and/or to make some parametric assumptions on the PDFs [20]. While these solutions may be satisfactory in some cases, we will discuss in Section 1.2 why they are inappropriate for tracking.

1.2 High-dimensional feature space

The combination of color and geometry proved to be efficient for tracking. In some works, spatial information has been added by means of a Gaussian weighting of the samples according to their distance to the center of the ROI [16, 38]. This weighting can be seen as a radial layout constraint. This approach has the advantage not to add any dimension to the feature space. However, it lacks generality. Geometry can instead be added directly to the radiometric vector (or any other feature vector), *e.g.*, in the form of the Cartesian coordinates of the pixels of the ROI [20]. Independence between color and geometry cannot be assumed in order to avoid to manipulate high-dimensional PDFs. Indeed, geometry alone, seen as a random variable conditionally to the ROI, follows a uniform distribution regardless of the ROI and, therefore, brings no information. While considering color and geometry jointly, simplification can still be achieved by approximating the PDFs with parametric laws [20]. Nevertheless, fully data-driven nonparametric PDF estimation was advantageously applied to segmentation [2, 30, 28] and this is the approach we propose to follow.

We also propose to extend the color and geometry feature space with the gradient of the luminance and patches of the luminance. The former was motivated by the fact that such a gradient has proved to increase accuracy in another motion-related task: optical flow computation [9]. The latter was motivated by studies or works such as [32, 11, 14]. Finally, we suggest to use the k-th nearest neighbor (kNN) framework in order to be able to handle the components of these high-dimensional feature vectors jointly and to work in a locally adaptive manner in the feature space, thus avoiding under or oversmoothing in processing the data set.

The following development applies to feature spaces of arbitrary dimension. In practice, though, the experiments that were performed to test the proposed method involve features of dimension 5, 7, or 13. Whether these dimensions can be considered as high is, in our opinion, mostly a matter of context. First, it is relative to the number of samples available. In tracking, this number can be rather small since it is given by the size of the (user-selected) ROI. Second, it depends on the purpose the samples are to be used for. When it comes to estimate statistical characteristics (entropies, divergences...), the denomination of high dimensional space makes sense since classical approaches already show their limits when the dimension gets higher than 2 or 3 (see Appendix A).

1.3 Similarity measure

Although kNN PDF estimators were proposed a long time ago [22, 34], they did not received much attention since they were known to be biased [42, 39]. Recently though, corrective terms have been derived to cancel the bias and led to consistent kNN-based statistical measures such as entropy [31, 26]. Moreover, even if the kNN PDF estimator is only adapted to high dimensions [42], the resulting entropy estimator appears to be accurate in both low and high dimensions.

In this context, the Kullback-Leibler divergence between high-dimensional PDFs will be suggested as a similarity measure for tracking. Although this measure has already been used for tracking [20], here the divergence will be expressed non-parametrically, making no assumptions, and directly from the samples, *i.e.* without explicit estimation of the underlying PDFs. This divergence estimator being well-adapted to high dimensions, it can be used in an extended radiometric/geometric feature spaces [6].

1.4 Paper organization and notations

The paper is organized as follows: Section 2 first provides some notations and general comments and then presents the kNN-based expressions of differential entropy, cross-entropy, and Kullback-Leibler divergence; Section 3 motivates the choice of geometry handling; Section 4 details the ROI tracking algorithm; Finally, Sections 5, 6, and 7 provide some results and comments for several standard sequences.

Please note that on-the-fly notations (used temporarily in a specific section) will be introduced by ":=". Also note that a statistical measure $\mathcal{M}$ function of a PDF f_U (*e.g.*, entropy) might appear as $\mathcal{M}(f_U)$ or $\mathcal{M}(U)$, where U is a set of samples drawn from f_U, whether it refers to the definition of the measure or a sample-based approximation of it.

2 Similarity measure between ROIs

2.1 Definition

Let I_{ref} be the reference frame in which the ROI domain Ω is (user-)defined and let I_{tgt} be the target frame in which the region which best matches this reference ROI (in terms of a given similarity measure) is to be searched for. Assume Ω is sampled on a, *e.g.*, Cartesian grid. At each grid node i, suppose a feature vector of $\mathbb{R}^d$ describing the frame locally at i can be built. For convenience, the set of grid nodes will also be denoted by Ω. Given the statistical approach chosen in Sections 1.2 and 1.3, the region search mentioned above amounts to finding the geometric transformation Φ such that

$$\Phi = \arg\min_{\varphi} \mathfrak{D}_{\text{KL}}(f_{T_\varphi}, f_R) \tag{1}$$

where $\mathfrak{D}_{\text{KL}}$ is the Kullback-Leibler divergence (or information gain) and f_R, resp. f_T, is the PDF which generated the reference feature samples $\{R(i), i \in \Omega\}$ in I_{ref}, resp. the target feature samples $\{T_\varphi(i), i \in \Omega\}$ in I_{tgt}. Whenever appropriate, U will be used as a generic notation for either R or T_φ.

The discussion below provides a motivation for choosing the order of the arguments of the divergence[1] (apparently not detailed in [20]).

Let us reformulate the problem in the following way: f_R is a reference PDF and the best Gaussian approximation $f^G_{T_\varphi}$ of it must be found. Minimizing $\mathfrak{D}_{\text{KL}}(f^G_{T_\varphi}, f_R)$ leads to a so-called zero-forcing solution [36]: wherever f_R is close to zero, the solution is strongly encouraged to be close to zero as well. As a consequence, $f^G_{T_\varphi}$ "focuses" on the dominant mode of f_R, thus underestimating the variance of f_R. This solution is also called exclusive since it can exclude some parts of f_R. Minimizing $\mathfrak{D}_{\text{KL}}(f_R, f^G_{T_\varphi})$ leads to a so-called zero-avoiding solution [36]: the solution is encouraged to cover the whole support of f_R. As a consequence, $f^G_{T_\varphi}$ usually overestimates the variance of f_R (see Fig. 1). Various works proposed symmetric versions of the Kullback-Leibler divergence, *e.g.*, J-divergence and Jensen-Shannon divergence [33]. Nevertheless, for tracking, $\mathfrak{D}_{\text{KL}}(f_{T_\varphi}, f_R)$ seems to be the appropriate choice. Indeed, f_{T_φ} can never be identical to f_R due to noise, occlusion, motion blur, and the fact that a frame is a projection onto a two-dimensional plane of a three-dimensional scene. However, both should have the same main modes if they correspond to the same object. Thus, the zero-forcing divergence enforces a relevant behavior in trying to "align" the main modes of the PDFs. By the way, it follows the same philosophy as the Bhattacharya distance, a measure widely used for tracking since a Mean-Shift-based implementation has been proposed [16].

[1] Remember that the Kullback-Leibler divergence is not symmetric.

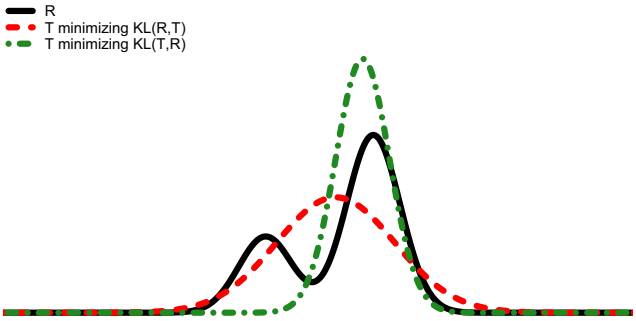


Figure 1: Order of the arguments of the Kullback-Leibler divergence: zero-forcing and zero-avoiding solutions (Image courtesy of Pierre Dangauthier, E-Motion project, INRIA Rhône Alpes/LIG, Grenoble, France).

As a reminder, the Kullback-Leibler divergence is equal to

$$\mathfrak{D}_{\text{KL}}(f_{T_\varphi}, f_R) = \int_{\mathbb{R}^d} f_{T_\varphi}(t) \ \log\left(\frac{f_{T_\varphi}(t)}{f_R(t)}\right) \ \mathrm{d}t \tag{2}$$

$$= H^{\times}(f_{T_\varphi}, f_R) - H(f_{T_\varphi}) \tag{3}$$

where H is the differential entropy

$$H(f_U) = -\int_{\mathbb{R}^d} f_U(t) \ \log f_U(t) \ \mathrm{d}t \tag{4}$$

and $H^{\times}$ is the cross entropy

$$H^{\times}(f_U, f_V) = -\int_{\mathbb{R}^d} f_U(t) \ \log f_V(t) \ \mathrm{d}t. \tag{5}$$

2.2 k-th nearest neighbor (kNN) estimation

○ Kernel-based approaches

Since PDFs are a central element of the proposed method, let us first discuss PDF estimation.

Kernel-based methods for PDF estimation make no assumption about the actual PDF. Consequently, the estimated PDF cannot be described in terms of a small set of parameters, as opposed to, *e.g.*, a Gaussian distribution defined by its mean and variance. These methods are therefore qualified as non-parametric. These estimators have the following general expression

$$f_U(t) = \frac{1}{|U|} \sum_{s \in U} K_{U,s,t}(t - s) \tag{6}$$

where $K_{U,s,t}$ is a multivariate kernel which bandwidth is a function of U, s, and t [42] and $|U|$ is the cardinality of the sample set U. Three cases can be distinguished

- $K_{U,s,t} = K_\sigma$ where the bandwidth σ is a constant. This is the Parzen approach. The choice of bandwidth σ is critical [41, 40].

  For a uniform kernel, estimator (6) approximates the density at t with the relative number of samples $k(t)/|U|$ falling into the open ball of volume v_σ centered on t

  $$f_U(t) = \frac{k(t)}{v_\sigma \ |U|}. \tag{7}$$

  Unfortunately, this kind of estimation provides a value too large when the underlying PDF has several modes. More generally, the Parzen method suffers from what is informally called the curse of dimensionality. As the dimension of the data space increases, the space sampling gets sparser. Therefore, less samples fall into the Parzen windows centered on each sample, making the PDF estimation less reliable. Dilating the Parzen window does not solve this problem since it leads to over-smoothing the PDF. In brief, this method cannot adapt to the local sample density.

- $K_{U,s,t} = K_{U,s}$. This is the sample point approach [42, 17]. One bandwidth is chosen per sample s of U. Although it allows to adapt to the local sample density, the following kNN framework was preferred since it leads to interesting statistical estimators such as the Kullback-Leibler divergence used in this paper.

- $K_{U,s,t} = K_{U,t}$. This is the balloon approach [34, 39]. The bandwidth is determined at each PDF estimation as a function of t. In the kNN framework, the bandwidth is defined by the distance to the k-th nearest neighbor of t among the samples of U. For a uniform kernel, estimator (6) reads [23][2]

  $$f_U(t) = \frac{k}{\rho_k^d(t) \ v_d \ |U|} \tag{8}$$

  where $\rho_k^d(t) \ v_d$ is the volume of the open ball centered on t with a radius of $\rho_k(t)$ equal to the distance to the k-th nearest neighbor of t in U excluding the sample located at t if any, and v_d is the volume of the unit ball in $\mathbb{R}^d$.

◦ First approximation of the divergence

The entropy (4) can be approximated by the Ahmad-Lin estimator [1]

$$H_{\mathrm{AL}}(U) = -\frac{1}{|U|} \sum_{s \in U} \log p_U(s) \tag{9}$$

where p_U is the Parzen estimation (6) of the actual PDF[3]. Approximation (9) converges in mean to the differential entropy of U.

The kNN PDF estimation (8) is biased and does not respect the fundamental PDF property of integrating to one. Nevertheless, these flaws get less critical as the dimensionality increases and the estimator has better overall performances in high dimensions than fixed bandwidth estimators [42]. Let plug (8) into the Ahmad-Lin entropy estimation (9)

$$H_{\mathrm{AL}}(U) \overset{\text{kNN}}{=} -\frac{1}{|U|} \sum_{s \in U} \log \frac{k}{\rho_k^d(U,s) \ v_d \ |U|} \tag{10}$$

$$= \log \frac{v_d \ |U|}{k} + \frac{d}{|U|} \sum_{s \in U} \log \rho_k(U,s). \tag{11}$$

Moreover, the cross entropy (5) is equal to

$$H^\times(f_U, f_V) = \mathrm{E}_U[-\log f_V] \tag{12}$$

$$\simeq -\frac{1}{|U|} \sum_{s \in U} \log f_V(s). \tag{13}$$

Again, plugging the kNN PDF expression of f_V into (13) leads to

$$H^\times(U,V) \overset{\text{kNN}}{=} \log \frac{v_d \ |V|}{k} + \frac{d}{|U|} \sum_{s \in U} \log \rho_k(V,s). \tag{14}$$

Subtracting (11) from (14), the following Kullback-Leibler approximation is obtained

$$\mathfrak{D}_{\mathrm{KL}}(T_\varphi, R) \overset{\text{kNN}}{=} \log \frac{|R|}{|T_\varphi|} + \frac{d}{|T_\varphi|} \sum_{s \in T_\varphi} \log \frac{\rho_k(R,s)}{\rho_k(T_\varphi,s)}. \tag{15}$$

Actually, this estimator has a slight bias. Nevertheless, the above development can help understanding the philosophy of the following, unbiased version.

◦ Unbiased version

Since the Kullback-Leibler divergence can be expressed as the difference between a cross entropy and an entropy, let us first present unbiased estimators of these quantities in the kNN framework.

Entropy. A consistent and unbiased entropy estimator was proposed for $k = 1$ [31]. This work was extended to $k > 1$ with a proof of consistency under weak conditions on the underlying PDF [26]

$$H_{\mathrm{kNN}}(U) = \log(v_d(|U|-1)) - \psi(k) + \frac{d}{|U|} \sum_{s \in U} \log \rho_k(U,s) \tag{16}$$

where v_d is the volume of the unit ball in $\mathbb{R}^d$, $|U|$ is the cardinality of the sample set U, ψ is the digamma

[2] See page 268.

[3] Note that an entropy estimation following the same spirit has been proposed more recently [43].

function Γ'/Γ, and $\rho_k(U,s)$ is the distance to the k-th nearest neighbor of s in U excluding the sample located at s if any. Informally, the main term in estimate (16) is equal to the mean of the log-distances to the k-th nearest neighbor of each sample. Note that (16) does not depend on the PDF f_U.

While the kNN PDF estimator is competitive in high dimensions only, the entropy estimator is accurate even in the univariate case [26]. Moreover, the choice of k appears to be much less crucial than the choice of σ in the Parzen method (see Appendix A). Actually, when the kNN approach is used for parameter estimation [8] (see Eq. (1)), k must be greater than the number of parameters, it must tend toward infinity when $|U|$ tends toward infinity, and such that $k/|U|$ tends toward zero when $|U|$ tends toward infinity. An admissible choice is $k = \sqrt{|U|}$.

Note that an estimate of the Rényi entropy using a related graph-based kNN framework has also been proposed for learning [18].

Cross entropy. Similarly, the cross entropy (also called relative entropy or likelihood) of two sample sets R and T_φ can be approximated by [31]

$$H^{\times}_{\mathrm{kNN}}(U,V) = \log(v_d|V|) - \psi(k) + \frac{d}{|U|}\sum_{s\in U}\log\rho_k(V,s). \tag{17}$$

Note again that estimator (17) does not depend on any PDF and that its main term is the mean of the log-distances to the k-th nearest neighbor among the samples of R of each sample of T_φ. Since a sample s of T_φ does not belong to R, the search for the k-th nearest neighbor *excluding* s itself does not in fact exclude any sample of R. This is why $|R|$ appears in (17) whereas $|T_\varphi|-1$ appears in (16).

Divergence. The Kullback-Leibler divergence can then be approximated in the kNN framework, directly from the sample sets R and T_φ, using the entropy and cross entropy estimators (16) and (17), resp.,

$$\mathfrak{D}_{\mathrm{KL}}(T_\varphi,R) \overset{\mathrm{kNN}}{=} H^{\times}_{\mathrm{kNN}}(T_\varphi,R) - H_{\mathrm{kNN}}(T_\varphi) \tag{18}$$

$$= \log\frac{|R|}{|T_\varphi|-1} + \frac{d}{|T_\varphi|}\sum_{s\in T_\varphi}\log\frac{\rho_k(R,s)}{\rho_k(T_\varphi,s)}. \tag{19}$$

It has been proven that this estimator is consistent and asymptotically unbiased [26, 31].

Remark about the biased version. Note that (19) only differs from (15) by $\log(|T_\varphi|/|T_\varphi-1|)$ in absolute value and that this difference tends toward zero when the number of target samples $|T_\varphi|$ tends toward infinity. Actually, concerning entropy and cross entropy, a similar remark can be made. Besides the term $|U|-1$ in (16) instead of $|U|$ in (11) (corresponding to the bias just mentioned about the divergence), the entropy estimators (11) and (16), and the cross entropy estimators (14) and (17) only differ by $\log(k)-\psi(k)$ in absolute value. Functions ψ being very close to log, this difference is also not very significant (see Table 1).

3 Feature space: handling geometry and radiometry

As noted earlier, the feature vectors combine radiometry and geometry. Radiometry allows to check if the ROI and the target region have similar colors and geometry allows to check with a given degree of strictness if these colors appear at the same location in the regions. For comparison purposes, Sections 3.1, 3.2, and 3.3 describe three levels of strictness. Let us assume that R and T_φ only contain radiometric information.

3.1 Geometry-free similarity measures

Classically, the similarity measure between the ROI and the target region can be a distance between color histograms or, similarly, PDFs. The knowledge of where a given color was present within the region is lost. For example, let us mention the Bhattacharya distance [16, 38]

$$\mathfrak{D}_{\mathrm{BHA}}(f_{T_\varphi},f_R) = \int_{\mathbb{R}^d}\sqrt{f_R(t)\ f_{T_\varphi}(t)}\ \mathrm{d}t \tag{20}$$

where d is equal to three if all color components are used. The Kullback-Leibler divergence on geometry-free PDFs will also be tested in Section 5.

Not accounting for the knowledge of where a given color was present in the region allows to be more flexible regarding the geometric transformation φ between the ROI and the target region. However, it increases the number of potential matches and then the risk for the tracking to fail after a few frames. This can be avoided by using a geometry-aware similarity measure.

3.2 Similarity measures with strict geometry

Geometry can be involved by means of a motion model (*i.e.*, a constraint on φ) used to compute a pointwise residual between the ROI and a candidate region. A function of the residual can serve as a similarity measure: classically, the SSD or functions used in robust estimation [5] such as the SAD. The geometric constraint being strictly defined by the motion model, these measures might be less efficient if the model is not coherent with the actual motion. Indeed, this might generate too many outliers in the residual, including in the

Table 1: Bias of the entropy estimator (11) and the cross entropy estimator (14) as a function of k.

Value of k	3	4	5	10	20	30	40
$\log(k)$	1.09	1.39	1.61	2.30	2.99	3.40	3.69
$\log(k)-\psi(k)$	0.18	0.13	0.10	0.05	0.03	0.02	0.01

framework of robust estimation. Moreover, even if the model is globally coherent with the actual motion, the choice of the function of the residual is implicitly linked to an assumption on the PDF of the residual, *e.g.*, Gaussian for SSD or Laplacian for SAD. This might not be valid in case of occlusion for example.

To fix the ideas, let us assume that $|T_\varphi| = |R|$ and let us define the following notations

$$\mathfrak{D}_{\text{SSD}}(T_\varphi, R) = \sum_{i\in\Omega} (T_\varphi(i) - R(i))^2 \tag{21}$$

and

$$\mathfrak{D}_{\text{SAD}}(T_\varphi, R) = \sum_{i\in\Omega} \phi(T_\varphi(i) - R(i)) \tag{22}$$

where ϕ can be either the absolute value or a smooth approximation of it, *e.g.*, $\phi(x) = \sqrt{x^2+\epsilon^2} - \epsilon$ [44].

3.3 Similarity measures with soft geometry

The geometric constraint can be soften, *e.g.*, by cascading a strict geometry approach and a radiometric approach [3] or, as proposed here, by adding geometry to the PDF-based approach presented in Section 3.1, *i.e.*, by defining a joint radiometric/geometric PDF [20, 6]. Formally, the PDF f_U corresponding to the sample set $\{U(i), i \in \Omega\}$ is replaced with the PDF $f_{U,i}$ corresponding to the sample set $\{(U(i), i), i \in \Omega\}$. Therefore, the color+geometry feature space is $\mathbb{R}^5$. In general, i can be any couple of independent spatial coordinates. For the ROI tracking application presented here, *normalized* Cartesian coordinates (x, y) were chosen: these coordinates are relative to the center of the bounding box of the ROI (*i.e.*, $(x, y) = (0, 0)$ at the center of the bounding box) and $\max(\max(|x|), \max(|y|)) = 1$ among the points of the ROI. Because geometry and radiometry are not comparable data, it might be useful or even necessary to weight one relatively to the other. It was decided to multiply the normalized coordinates by a spatial weight δ, resulting in $\max(\max(|x|), \max(|y|))$ being equal to δ.

3.4 Enrichment of the radiometric constraint

As mentioned earlier, the proposed kNN framework is valid for any feature space dimension d. In Section 5, it will be clear that color and geometry as combined in Section 3.3 can provide enough information even in challenging situations. Yet, if it accounts for the correlation between a color and its location of appearance, it does not account for the correlation between the colors of neighboring pixels. This could be done by involving, *e.g.*, the color gradient or patches [32, 14] (see Section 6). The influence of the chosen feature space, involving geometry and radiometry in several ways, is illustrated in Fig. 2. In our experiments, the following cases will be tested:

- $U(i) = I(i)$;
- $U(i) = (I(i), \gamma \nabla I_Y(i))$;
- and $U(i) = (\text{Patch}_{3\times3}(I_Y(i)), I_U(i), I_V(i))$;

where I_Y is the luminance component of I, (I_U, I_V) are the chrominance components, $\text{Patch}_{3\times3}(I(i))$ is a 3×3-patch of I centered at i, and γ is a constant.

4 Tracking algorithm

4.1 The main steps

We propose to perform tracking by minimizing the kNN Kullback-Leibler divergence (19) with respect to φ, or actually a set of parameters defining φ. The chosen motion model is "translation+scaling"

$$\varphi(i) = i + M(i)\, p \tag{23}$$

$$= \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} x & 1 & 0 \\ y & 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha - 1 \\ u \\ v \end{bmatrix} \tag{24}$$

where α is the scaling factor and (u, v) is the translation. The main steps of the tracking algorithm are presented in Table 2. The tracking result is represented by the set $\{\varphi_{\text{tgt}}\}$.

4.2 Series of minimizations

The minimization of (19) with respect to $\varphi = (\alpha, u, v)$ can be performed by a series of minimizations in (u, v) at α fixed, as illustrated in Table 2. This decoupling allows to confine α to a reasonable interval, *e.g.*, $[0.98, 1.02]$. The minimizations in (u, v) can be achieved by a gradient descent setting the α-component of gradient (25) to zero. For computational considerations, they can instead be performed

Figure 2: Influence of the feature space. The pixels in green are the 500 nearest neighbors of the pixel marked with the white cross. The feature space is composed of *(in lexicographical order)* geometry only; grayscale intensity only; grayscale intensity and geometry; grayscale intensity, geometry, and gradient; and 3×3-patch and geometry. When the gradient is added to grayscale intensity and geometry, most of the neighboring pixels are located on edges with a similar gradient norm and orientation (besides having a similar color and not being to far away in the image plane) to the pixel of reference.

Table 2: Tracking algorithm.

- Set the parameters
  - Neighboring order: $k \overset{e.g.}{\leftarrow} 3$
  - Spatial weight: $\delta \overset{e.g.}{\leftarrow} 1$
  - Scaling factors: $\lambda \overset{e.g.}{\leftarrow} \{0.98, 0.99, 1, 1.01, 1.02\}$
  - Radiometric function: $U(i) \overset{e.g.}{=} I(i)$
- Manually select a ROI Ω in the reference frame I_{ref}
  1. Let $i_R = (x_R, y_R)$ be the normalized Cartesian coordinate system relative to Ω
     Perform either 2 or 3 depending on the minimization strategy (see below)
  2. *Either:* Set $R_\alpha = \{(I_{\text{ref}}(i_R), \alpha\delta i_R),\ i_R \in \Omega\}$ for all $\alpha \in \lambda$
  3. *Or:* Set $R = \{(I_{\text{ref}}(i_R), \delta i_R),\ i_R \in \Omega\}$
- Let φ be the triplet (α, u, v) equal to $(1, 0, 0)$ initially
- For each remaining frame I_{tgt} taken sequentially
  1. Let $i_T = (x_T, y_T)$ be the normalized Cartesian coordinate system relative to $\varphi(\Omega)$
     Perform minimization using either strategy 2 or strategy 3
  2. *Either:* Perform a series of minimizations as follows
     (a) For each $\beta \in \lambda$
     - Determine the translation (m, n) such that
     $$(m, n) = \arg\min_{(a,b)} \mathfrak{D}_{\text{KL}}(T_{(a,b)}, R_\beta)$$
     where $T_{(a,b)} = \{I_{\text{tgt}}(i_T + (a, b)), \delta i_T),\ i_T \in \varphi(\Omega)\}$ (see Section 4.2)
     - Let $\mathfrak{D}_\beta$ be equal to $\mathfrak{D}_{\text{KL}}(T_{(m,n)}, R_\beta)$

     (b) Determine the triplet $(\tilde{\beta}, \tilde{m}, \tilde{n})$ that gave the lowest $\mathfrak{D}_\beta$ among the $|\lambda|$ loops of 2a
  3. *Or:* Perform a gradient descent in (α, u, v) (see Section 4.3) to determine the triplet $(\tilde{\beta}, \tilde{m}, \tilde{n})$ that minimizes $\mathfrak{D}_{\text{KL}}(T_{(m,n)}, R_\beta)$ where R_β is obtained by multiplying the geometry stored in R by β
  4. $\varphi = (\alpha, u, v) \leftarrow (\alpha\ \tilde{\beta}, u + \tilde{m}, v + \tilde{n})$
  5. $\varphi_{\text{tgt}} \leftarrow \varphi$

using a suboptimal search procedure such as the diamond search [46], thus following the approach for block matching of standard video coders. Naturally, more sophisticated search techniques such as particle filters [38][4], also known as sequential Monte Carlo methods, can be used.

4.3 Mean-Shift-based gradient descent

Estimation (19) being defined in the kNN framework, it is not differentiable. Alternatively, one could think of using the Parzen formulation of the PDFs and the Mean-Shift approximation to determine the derivative of the Kullback-Leibler divergence (see Appendix B) and then evaluating the derivative using the kNN framework (see Appendix C)

$$\nabla_\varphi \mathfrak{D}_{\text{KL}}(T_\varphi, R) = -\frac{1}{k\ |T_\varphi|} \sum_{s \in T_\varphi} \mathcal{D}_s(T_\varphi) \left(\frac{d+2}{\rho_k^2(R, s)} \sum_{t \in W_{\rho_k(R,s)}} (t - s) - \frac{d+2}{\rho_k^2(T_\varphi, s)} \sum_{t \in W_{\rho_k(T_\varphi, s)}} (t - s) - \sum_{\substack{t \in T_\varphi \\ |t-s| = \rho_k(T_\varphi, t)}} \frac{t - s}{\rho_k(T_\varphi, t)} \right) \tag{25}$$

where $\mathcal{D}_s(T_\varphi)$ is a $3 \times d$-matrix involving frame gradients and $W_{\rho_k(\cdot,s)}$ is a window of radius $\rho_k(\cdot, s)$ centered at sample s. (The definitions of all the involved terms are given in the developments presented in Appendices B and C which lead to derivative (25).) As a consequence, the ROI tracking could be solved by gradient descent in the space of the parameters (α, u, v). However, the sensitivity of the similarity measure with respect to the scaling α is much higher than the sensitivity with respect to translation. In practice, this can lead to undesirable convergence behaviors such as finding a match in the target frame at a scale different from the scale of the reference ROI (*i.e.*, the reference could be matched to a region much larger or much smaller). Therefore, a procedure based on a series of minimizations might be preferable (see Section 4.2).

5 Experimental results - Part 1

5.1 Setup

The proposed kNN-based algorithm presented in Section 4 will be referred to as kNN-KL-G where KL stands for Kullback-Leibler and G stands for geometry.

[4] These methods are particularly efficient in case of total occlusion of the target on several frames.

It was compared to four other trackers: *(i)* a geometry-free version of the proposed method (kNN-KL), *(ii)* a version of the proposed method where the kNN expression (19) of the divergence was replaced with an estimation based on Parzen windowing[5] (Pz-KL-G), *(iii)* an SAD version of the algorithm described in Table 2 (*i.e.*, replacing the Kullback-Leibler divergence in step 2a by energy (22)), and *(iv)* a Mean-Shift-based tracker whose implementation is publicly available [15].

Note that in these comparisons, we focused on the pros and cons of the different similarity measures and their approximations. To try to avoid "corruption" of the results by other methodological aspects, we kept the tracking algorithm simple, purposely setting aside improvements such as reference update and motion prediction. Moreover, for a fair comparison between all these methods, the experimental setup of the above-mentioned Mean-Shift implementation was followed, namely, a rectangular ROI Ω (see Figs. 3, 4, and 6 for the dimensions) and a translation only motion φ (*i.e.*, $\lambda = \{1\}$) with a pixel resolution. The chosen radiometric space was YUV simply because the standard test sequences used in our experiments are available in this color space.

For the kNN-based methods, parameter k in (19) was chosen equal to 3, which satisfies the conditions mentioned after Eq. (16). An experimental study of the stability of the proposed method with respect to this parameter is presented in Section 5.7. The distance $\rho_k(U, s)$ to the k-th nearest neighbor of s in U was defined in the classical Euclidean sense. For its computation, we used a publicly available implementation [27].

The components of the feature vectors were normalized as follows: Y, U, and V were rescaled into the interval $[0, 1]$ and, as explained in Section 3.3, the coordinates (x, y) were rescaled into $[-1, 1]$, both in the ROI and the candidate regions, the origin being located at the center of the bounding box of the region. The spatial weighting δ was taken equal to 1.

The minimization in $\varphi = (u, v)$ was implemented using a suboptimal search procedure known as the diamond search [46]. Tracking was performed with I_{ref} being set to I_1 while I_{tgt} was successively equal to I_t, $t = 2, 3, 4 \ldots$ When searching for the ROI in frame I_t, the search area was empirically limited to ± 12 pixels horizontally and vertically around the position of the center of the ROI computed in frame I_{t-1}.

5.2 Partial occlusions

Sequence "Car" is an aerial car chase which is part of the VIVID tracking testbed [15]. It is composed of 640×480-frames. Tracking was performed on 150 consecutive frames (see Fig. 3). kNN-KL eventually lost the ROI and ended up tracking the second car which has colors similar to the ROI. This is probably due to the fact that it is based on radiometry only. Pz-KL-G also failed in tracking the first car. Mean-Shift performed quite well although the tracking shifted upward when occlusion occurred in order to avoid including

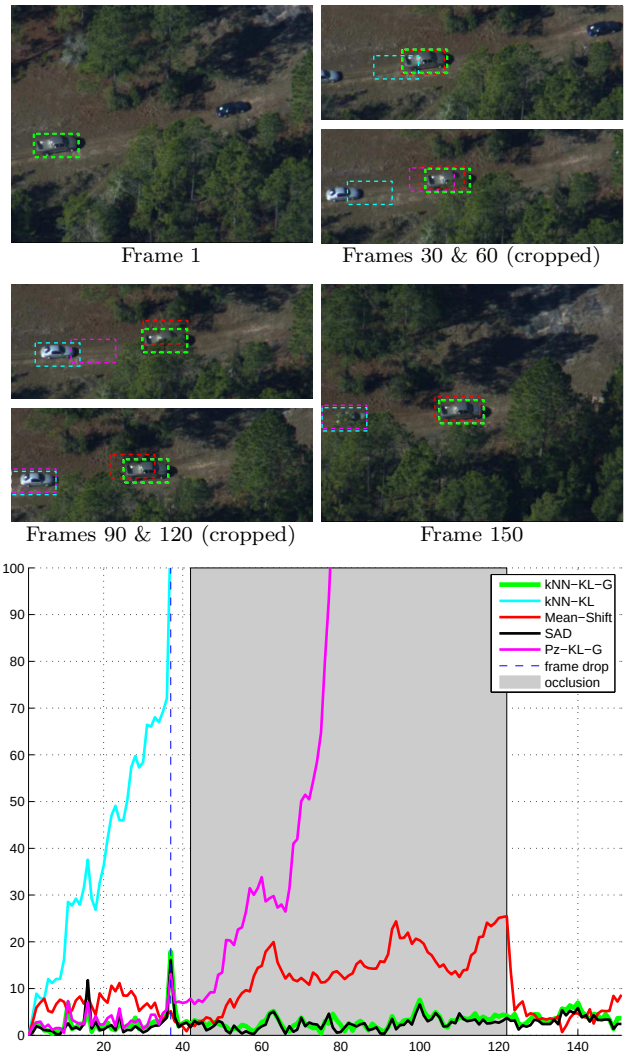


Figure 3: Tracking on sequence "Car" (frame indices are relative to the reference frame). kNN-KL-G (proposed method): green; kNN-KL: cyan; Pz-KL-G: pink; Mean-Shift: red; SAD: white on the frames and black in the diagram. There is a frame drop of several frames at frame 38 (vertical dashed line in the diagram) and the tracked car is partially occulted by trees from frame 42 to frame 122 (gray area in the diagram). The diagram represents the shift (in percent of the ROI diagonal length in pixel) with respect to a manually defined tracking as a function of the frame index. Ω: 95×47-rectangle.

[5] This Kullback-Leibler implementation is publicly available [29].

the green colors of the trees in the color PDF. Concerning SAD, the translation model being fairly well respected within the ROI, taking the pointwise residual makes sense while the use of the absolute value is robust to the outliers arising from the occlusion. As a consequence, the car was accurately tracked. Finally, kNN-KL-G also performed very well.

5.3 Variations of luminance

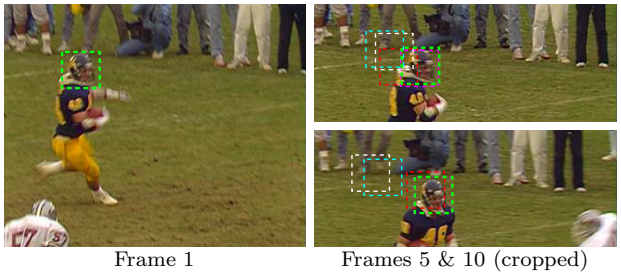
Sequence "Crew" is composed of 352×288-frames. Two faces were tracked on 80 consecutive frames (see Fig. 4). kNN-KL-G tracked the faces successfully. The other methods sooner or later lost the ROIs, apparently due to the variations of luminance. This is particularly obvious with *(i)* Mean-Shift which brutal changes in tracking shift match the camera flashes in frames 16, 20, and 61 for the face on the left, and in frames 1, 7, and 61 for the face on the right, and *(ii)* kNN-KL which tracking error seems to follow the curve of the average intensity.

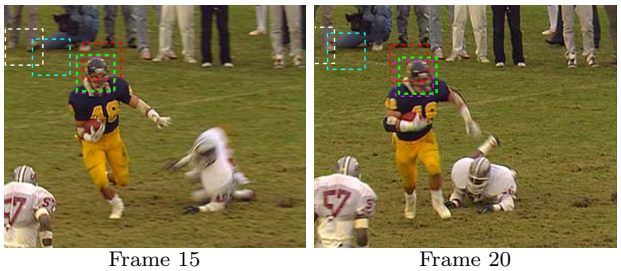
5.4 Noisy sequence

Sequence "Schnee" is composed of 768×576-frames. Two cars were tracked on 160 consecutive frames (see Fig. 5). This sequence can be considered noisy due to the snowflakes which fall rather densely. Despite this "Salt" noise, the two cars were accurately tracked by kNN-KL-G. SAD also performed quite well. The objects being small and rather homogeneous, their motion could be considered as a translation. Therefore the strict geometric constraint of SAD is not violated. Clearly, Mean-Shift was disturbed by the noise. The other two methods (kNN-KL and Pz-KL-G) worked pretty well for one car but failed for the other one. These methods have their similarity measure in common. However, only one involves geometry. This result appears difficult to interpret and might only be fortuitous.

5.5 Complex motion

Sequence "Football" is composed of 352×288-frames. Tracking was performed on 20 consecutive frames (see Fig. 6). Note that part of the public has colors similar to colors that can be found in the ROI. In some frames, this area of the public is right above the ROI. This is probably the reason why kNN-KL stayed stuck in this region. Moreover, as the player runs, he turns and almost faces the camera toward the end of the sequence. Therefore, the translation model is not appropriate. This can explain why SAD, which relies on a strict translation model, lost the ROI in the first frames. Mean-Shift succeeded to track the ROI approximately. However, it could not avoid being at-


Frame 1 Frames 5 & 10 (cropped)


Frame 15 Frame 20

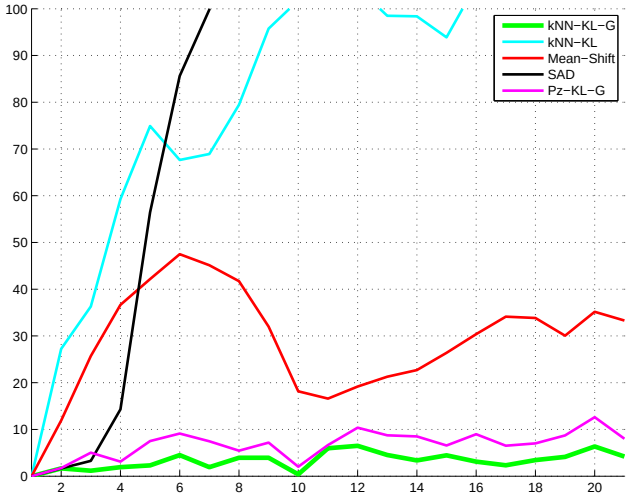


Figure 6: Tracking on sequence "Football" (frame indices are relative to the reference frame). kNN-KL-G (proposed method): green; kNN-KL: cyan; Pz-KL-G: pink; Mean-Shift: red; SAD: white on the frames and black in the diagram. This sequence is characterized by a fast motion generating motion blur. Moreover, the motion of the object of interest has a rotational component responsible for the disappearance of some areas and the exposure of others. The diagram represents the shift (in percent of the ROI diagonal) with respect to a manually defined tracking as a function of the frame index. Ω: 43×43-square.

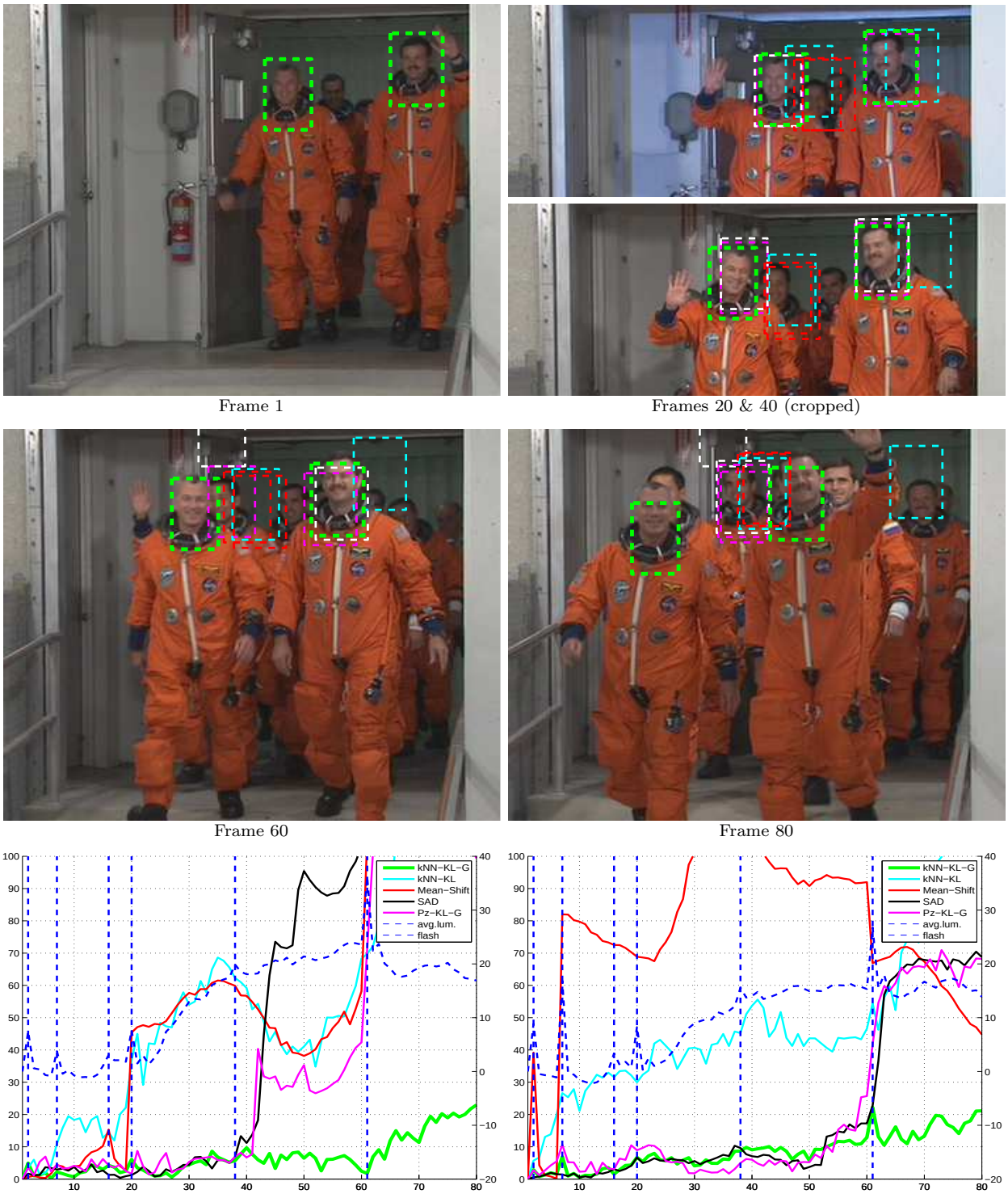


Figure 4: Tracking on sequence "Crew" (frame indices are relative to the reference frame). kNN-KL-G (proposed method): green; kNN-KL: cyan; Pz-KL-G: pink; Mean-Shift: red; SAD: white on the frames and black in the diagrams. There are two kinds of intensity changes in the sequence: a slight, continuous intensity increase as the crew walks out of a dark area, and some strong and brief intensity peaks due to camera flashes (vertical dashed lines in the diagrams). The diagrams represent the shift (in percent of the ROI diagonal length in pixel) with respect to manually defined trackings as a function of the frame index. The diagram on the top corresponds to the face on the left. The vertical axis on the right of each diagram corresponds to the blue dashed curves which represent the evolution of the average intensity (Y component) within the manually defined trackings. The average intensity in frame 1 is taken as the reference and the scale is in unit of intensity. Both the continuous intensity increase and the camera flashes are noticeable. Ω: 33×52-rectangle.

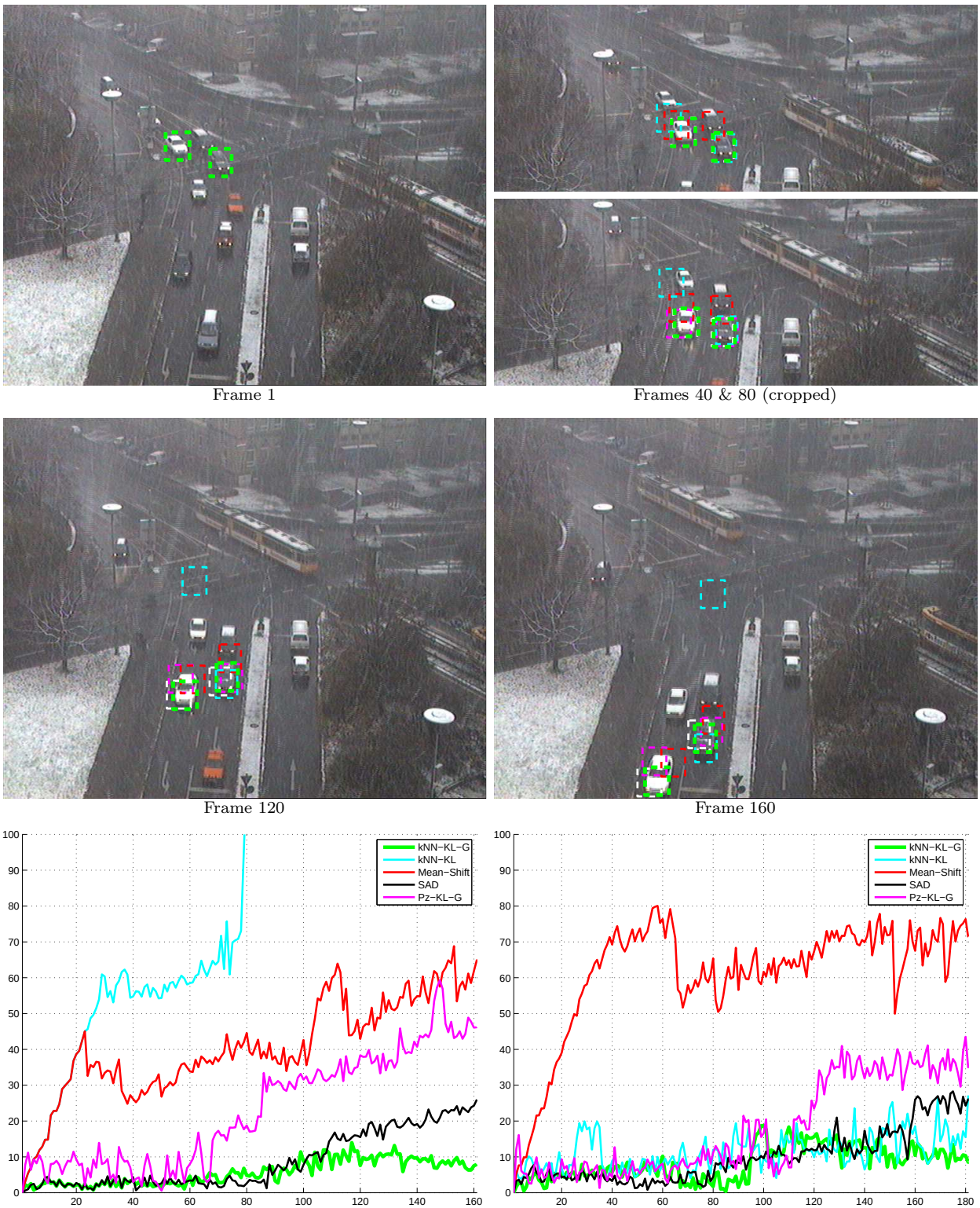


Figure 5: Tracking on sequence "Schnee" (frame indices are relative to the reference frame). kNN-KL-G (proposed method): green; kNN-KL: cyan; Pz-KL-G: pink; Mean-Shift: red; SAD: white on the frames and black in the diagrams. This sequence can be considered noisy due to the snowflakes. The diagrams represent the shift (in percent of the ROI diagonal) with respect to manually defined trackings as a function of the frame index. The diagram on the top corresponds to the car on the left. Ω: a 38×42-square for the car on the left and a 34×42-square for the car on the right.

Table 3: Summary of the comparisons on the four sequences "Car", "Crew", "Schnee", and "Football".

Fig.	Best results	Worst results
3	kNN-KL-G, SAD & (to some extent) Mean-Shift	kNN-KL & Pz-KL-G
4a	kNN-KL-G	kNN-KL & Mean-Shift
4b	kNN-KL-G & (to some extent) SAD & Pz-KL-G	kNN-KL & Mean-Shift
5a	kNN-KL-G & SAD	kNN-KL & Mean-Shift
5b	kNN-KL-G, SAD, kNN-KL & (to some extent) Pz-KL-G	Mean-Shift
6	kNN-KL-G & Pz-KL-G	kNN-KL & SAD

tracted by the public. The geometric constraint of kNN-KL-G and Pz-KL-G allowed to avoid being attracted by the public area (where the color spatial arrangement is different from that of the reference ROI) while being soft enough to deal with the mismatch between the translation model and the actual motion. The resulting trackings are accurate. (Nevertheless, kNN-KL-G performed better than Pz-KL-G, arguably because it relies on variable kernel bandwidth.)

To support these conclusions, the distance between the reference ROI and candidate regions in frame 20 was computed as a function of the translation parameters for SAD, kNN-KL, Pz-KL-G, and kNN-KL-G (see Fig. 7). The red spot at the center of the plane represents the correct motion. The SAD minimum is shifted as a result of the inappropriateness of the translation model between frame 1 and frame 20. kNN-KL has several local minima as there are several possible matches when accounting for radiometry only. By adding geometry, Pz-KL-G allows to find a unique minimum, although not at the right location. This is certainly due to the reduced accuracy of the Parzen-based estimator of the statistical measure in $\mathbb{R}^5$. Finally, kNN-KL-G has a minimum that matches the correct motion. Also note that the kNN-KL-G criterion seems strictly convex in a large window around the minimum. This property is interesting for the convergence of optimization algorithms (diamond search in our case).

5.6 Summary

The previous comparisons could be coarsely summarized by selecting the two or three best and worst methods for each of the four sequences (see Table.3). The conclusions that could be made are:

Table 4: Stability of kNN-KL-G with respect to k: average norm of the tracking shifts, norm of the sum of the shifts (both in percent of the ROI diagonal length in pixel), and orientation of the sum of the shifts (in degree) taking the result obtained with $k = 3$ as a reference.

Value of k	3	10	20	43= $\sqrt{\lvert\Omega\rvert}$
Avg norm	Ref.	0.46	1.69	2.65
Sum norm	Ref.	0.60	1.80	1.30
Sum angle	Ref.	118	-68	117

- kNN-KL almost always fails. This is a known effect of not taking geometry into account (see Section 3.1);
- Mean-Shift fails in most of our tests (variation of illuminance and noise) but can also perform quite well;
- SAD might represent a computationally efficient alternative to kNN-KL-G if accuracy is not a major requirement. Unfortunately, it can fail completely when the motion is complex (see Fig. 6) since it relies on a strict geometric constraint (see Section 3.2);
- The performance of Pz-KL-G ranges from reasonably good to totally unacceptable. It relies on the Parzen approach instead of the proposed kNN framework to estimate the chosen statistical measure and therefore allows to illustrate the expected advantages of kNN (see Section 2.2 and Appendix A);
- Finally, kNN-KL-G represents the best option in all cases.

5.7 Stability with respect to k

To evaluate the stability of kNN-KL-G with respect to the choice of parameter k, tracking was performed on sequence "Football" with various values of k that comply the conditions mentioned in Section 2.2. The tracking obtained for k equal to 3 was taken as a reference and the average shifts over the 20 frames resulting from using other values were measured (see Table 4). In each frame, the vectorial shift between the bounding box obtained for $k = 3$ and the bounding box obtained for another value of k was determined. The second line in Table 4) corresponds to the average of their norm. It tells us that, as k gets further away from the chosen value of reference, the solution of the tracking has also a tendency to shift away from the solution of reference. This is not surprising and, looking at the numbers, this behavior is not excessive. The third and fourth lines of

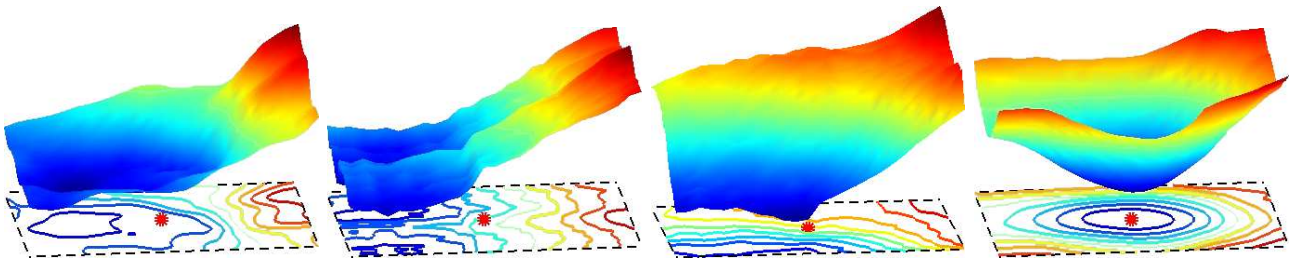


Figure 7: Distance between the reference ROI of sequence "Football" and candidate regions in frame 20 as a function of horizontal and vertical translations: *(in lexicographical order)* SAD, kNN-KL, Pz-KL-G, and kNN-KL-G (proposed method). The dashed box is a 12×12-square (same size as the search window). The red spot at its center represents the correct translation. The SAD and Pz-KL-G minima are shifted, kNN-KL has two local minima, and the minimum of kNN-KL-G seems accurate.

Table 4) correspond to, resp., the norm and the orientation of the sum of the the shift vectors in the successive frames. They tell us about the coherence of the shifts for the different values of k. Apparently, there is no such coherence, the summed shifts being of negligible norms with various orientations. In conclusion, as k increases, the tracking shift tends to oscillate more and more around the solution for $k = 3$, but confined in an acceptable range and without any obvious coherence. Therefore, the method appears to be quite stable with respect to k.

6 Experimental results - Part 2

6.1 Setup

From now on, the proposed method will be compared with variants of itself. Consequently, there are no constraints on the experimental setup and then scaling will be taken into account by choosing $\lambda = \{0.98, 0.99, 1, 1.01, 1.02\}$, and the gradient of the luminance and patches of the luminance will be optionally used as additional radiometric features.

The frames of the sequences already presented were available in the YUV color space. In the following, another sequence [21] will be used. It was available in the RGB space but will be converted to the YUV space before processing. The components of the feature vectors were normalized as follows: Y, U, and V, were rescaled into the interval $[0, 1]$, the gradient of the luminance Y (whenever used) was computed using the filter $[-1, 9, -45, 0, 45, -9, 1]/60$ and rescaled using $\gamma = 10$, and, as a reminder, the coordinates (x, y) were rescaled into $[-1, 1]$. The spatial coordinates were further modified by applying the spatial weighting δ for the target and the scaling $\alpha\delta$ for the reference, meaning that (x_R, y_R) actually belongs to the interval $[-\alpha\delta, \alpha\delta]^2$ and (x_T, y_T) belongs to $[-\delta, \delta]^2$.

The minimization in $\varphi = (\alpha, u, v)$ was either performed by a series of minimizations at α fixed (see Section 4.2) implemented using a suboptimal search procedure known as the diamond search [46], or by a gradient descent procedure: for stability, the gradient (25) was normalized such that the translation component has a norm equal to one and the scaling component is either 0.99 or 1.01. The former minimization strategy will be referred to as "Discrete search" and the latter one as "Gradient search".

The other aspects of the setup were identical to Section 5.1.

6.2 A first example with scaling

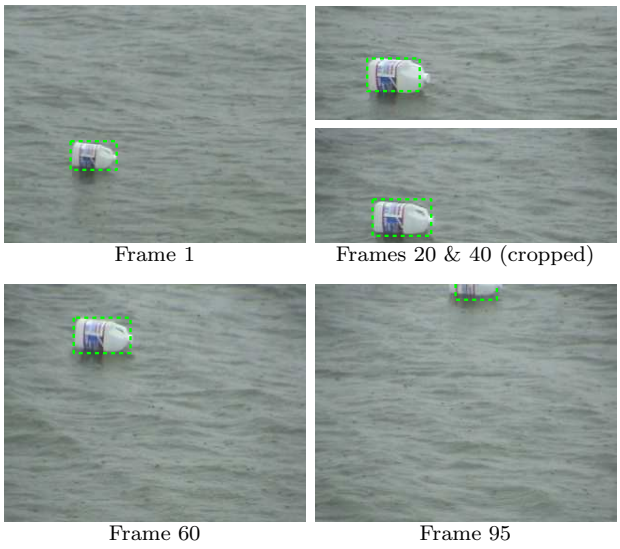
Sequence "WaterObject" is composed of 352 × 288-frames. Tracking was performed on 95 consecutive frames using the discrete search with $\delta = 1$ and radiometry limited to the color information (see Fig. 8).

6.3 Influence of δ

A tracking was performed on 60 consecutive frames of sequence "Crew" using the discrete search and two values of δ: 0.6 and 1. The radiometric information was limited to color (see Fig. 9). As expected, the spatial weighting has an influence on the tracking quality. Nevertheless, it is not dramatic since it seems to play mostly on the duration the tracking can be considered accurate rather than acting on the stability of the processing.

6.4 Gradient as an additional radiometric feature

A tracking was performed on 150 consecutive frames of sequence "Crew" using the discrete search and $\delta = 0.6$. The feature space was either color+geometry or color+gradient of the luminance+geometry (see Fig. 10). Clearly, the addition of the gradient information improved the tracking accuracy. As mentioned



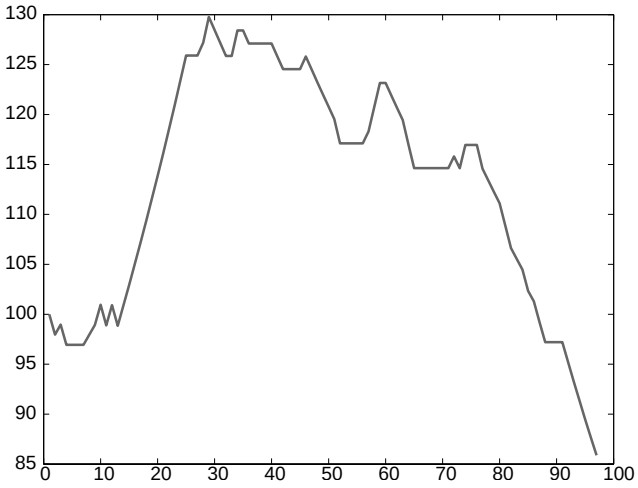


Figure 8: Tracking on sequence "WaterObject" (frame indices are relative to the reference frame). This sequence is characterized by zooms in and out. The diagram represents the scaling of the ROI (parameter α times 100) as a function of the frame index. To deduce the scaling in terms of area, the value must be divided by 100 and squared. At the highest point, the ROI is almost 1.7 times larger in area than Ω, a 48×28-square, and around 0.7 times smaller at the lowest point.

earlier, any other feature can be added without algorithm modifications (it only add terms to the Euclidean distance computation between the feature vectors during the kNN search). It does not necessarily mean that more features implies better tracking accuracy (see Section 6.6).

6.5 Discrete search vs. gradient search

Sequence "Poltergay" [21] is composed of 720×576-frames in JPEG format. Tracking was performed on 100 consecutive frames using either the discrete search or the gradient search with $\delta = 0.8$ and the feature space defined as color+gradient of the luminance+geometry (see Fig. 11). As expected, the discrete search performed better than the gradient search due to the presence of local minima which can mislead a gradient descent (the gradient search performed very decently, though). However, this is at the cost of a computational time 7 times higher.

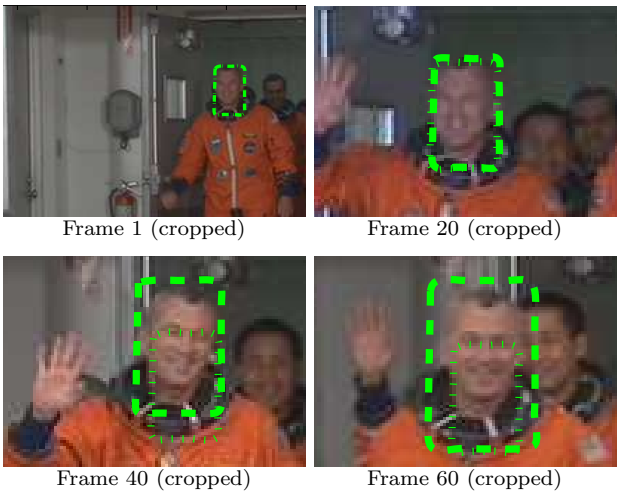


Figure 9: Tracking on sequence "Crew" (frame indices are relative to the reference frame). $\delta = 1$: dotted line; $\delta = 0.6$: dashed line. Ω: 33×52-rectangle.

6.6 Noisy sequences

A tracking was performed on 100 consecutive frames of two degraded versions of sequence "Poltergay" using the discrete search, $\delta = 0.8$ and the feature space being either *(i)* color+geometry, *(ii)* color+gradient of the luminance+geometry, or *(iii)* "patch 3×3 on Y"+U+V+geometry (space of dimension 13). The first degradation was obtained by compressing each frame at a very low rate using a JPEG2000 coder (see Fig. 12). The original frames in JPEG format have a size of around 32 kB. The JPEG2000 compression rate was chosen such that the size went down to 4kB. For the second degradation, each color channel of each frame was corrupted by a Gaussian noise of mean zero and variance equal to 9 (see Fig. 12).

The proposed method being independent of the ROI shape, the rectangular shape used so far for Ω was replaced with an ellipse with a bounding box of 63×101 pixels.

Both feature spaces *(i)* and *(ii)* dealt very well with the JPEG2000 artifacts (see Fig. 13). Therefore, feature space *(iii)* was not even considered. However, since the Gaussian noise largely corrupted the gradient of the frames, the color+gradient+geometry feature space did not allow to track the object. Although

Figure 12: Detail of frame 1 of sequence "Poltergay" *(in lexicographical order)* uncorrupted, corrupted by JPEG2000 compression artifacts, and corrupted by a Gaussian noise.

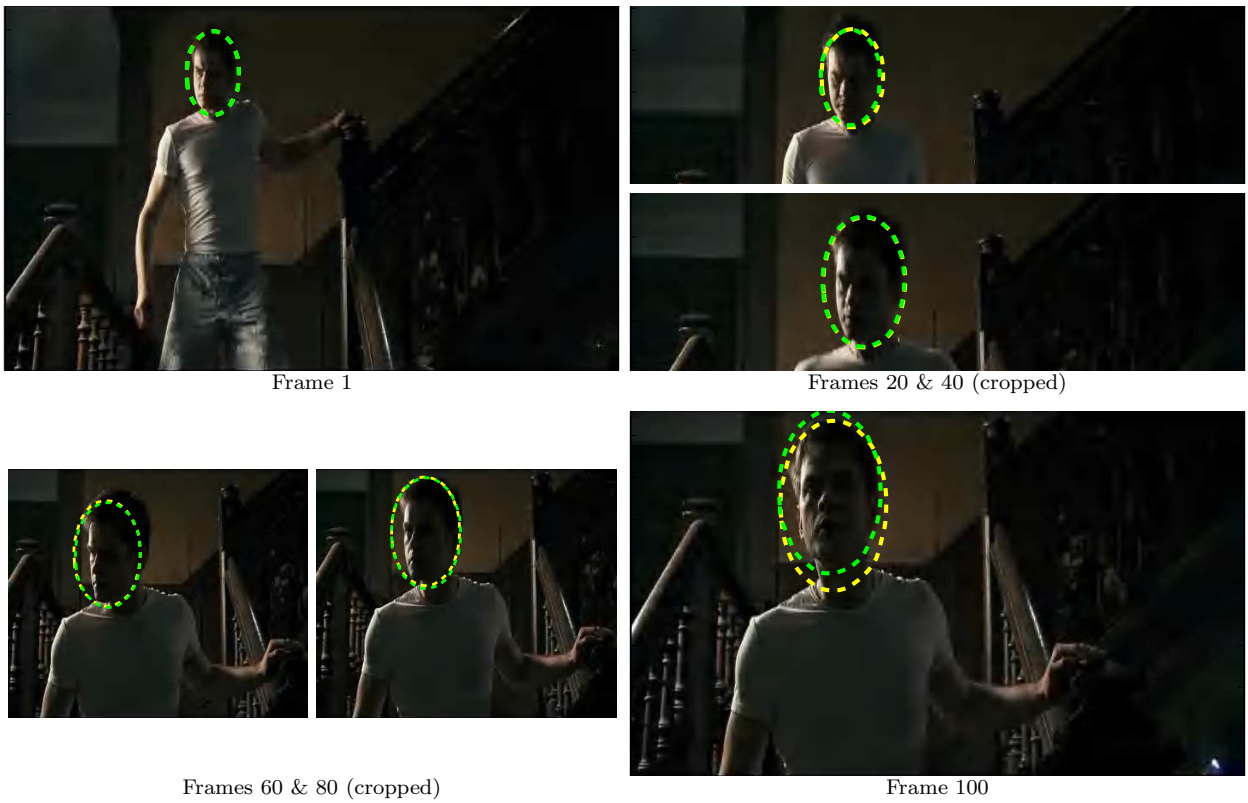
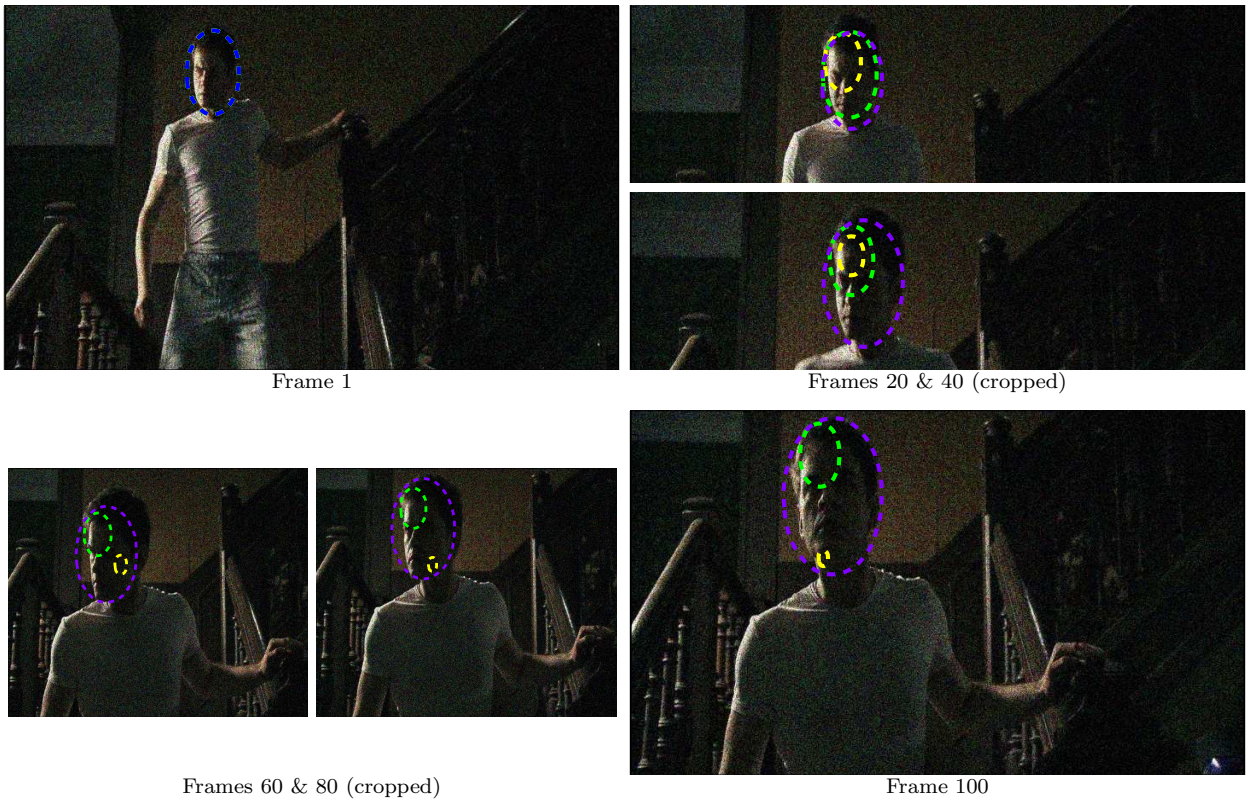


Figure 13: Tracking on sequence "Poltergay" in the presence of JPEG2000 compression artifacts (frame indices are relative to the reference frame). Without the gradient of the luminance (feature space *(i)*): green; with the gradient (feature space *(ii)*): yellow. Ω: ellipse with a bounding box of 63×101 pixels.



Frame 1 | Frames 20 & 40 (cropped)

Frames 60 & 80 (cropped) | Frame 100

Figure 14: Tracking on sequence "Poltergay" in the presence of Gaussian noise (frame indices are relative to the reference frame). Without the gradient of the luminance (feature space *(i)*): green; with the gradient (feature space *(ii)*): yellow; with the patches (feature space *(iii)*): blue. Ω: ellipse with a bounding box of 63×101 pixels.

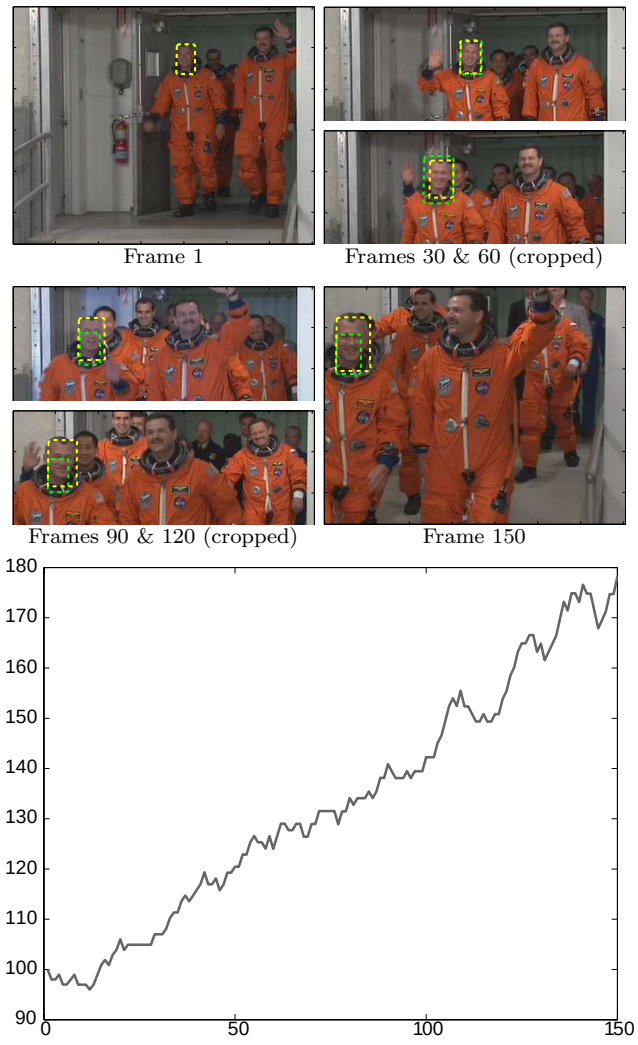


Figure 10: Tracking on sequence "Crew" (frame indices are relative to the reference frame). Without the gradient of the luminance: green; with the gradient: yellow. The diagram represents the scaling of the ROI (parameter α times 100) as a function of the frame index for the solution that used the gradient. To deduce the scaling in terms of area, the value must be divided by 100 and squared. At the highest point, the ROI is more than 3 times larger in area than Ω, a 33×52-rectangle.

not satisfying, the color+geometry space produced a more acceptable tracking (actually, it is even quite accurate until frame 22 (not shown in Fig. 14)). It is only when considering patches (feature space *(iii)*) that the tracking becomes fully accurate (see Fig. 14). This robustness to noise of patches is not surprising since they were used for denoising [11, 12].

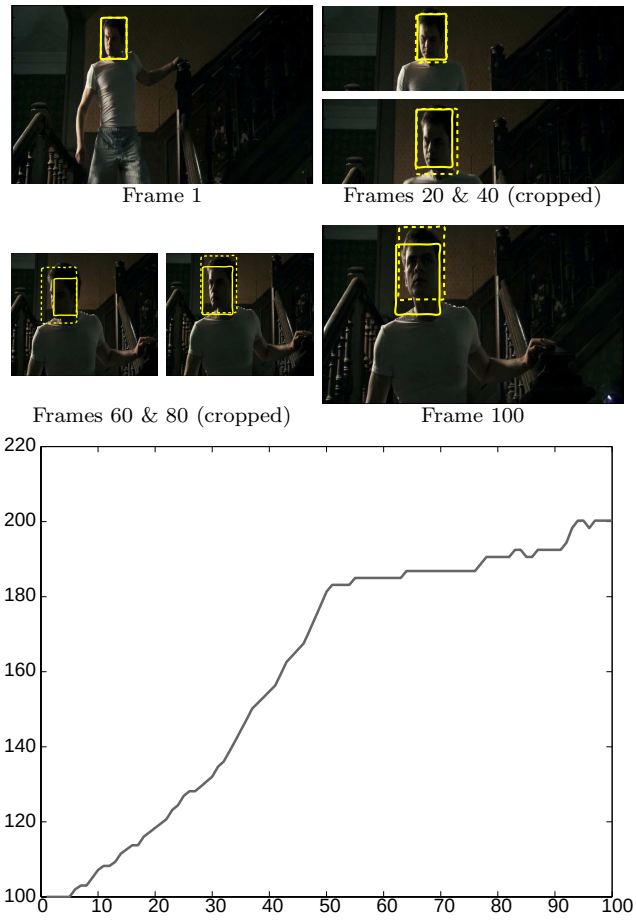


Figure 11: Tracking on sequence "Poltergay" (frame indices are relative to the reference frame). Discrete search: dashed line; gradient search: plain line. The diagram represents the scaling of the ROI (parameter α times 100) as a function of the frame index for the solution using the discrete search. To deduce the scaling in terms of area, the value must be divided by 100 and squared. At the highest point, the ROI is more than 4 times larger in area than Ω: 63×101-rectangle.

7 Conclusion and ongoing works

This paper presents a general framework for estimating high-dimensional statistical measures to perform ROI tracking. We focused on a measure derived from entropy as proofs of consistency and unbiasedness exist [26, 31].

The kNN-based PDF estimate has two advantages for dealing with high-dimensional data. First, it relies on a non-parametric approach which uses variable size kernels to adapt to the local density of samples. Second, it allows to derive expressions of PDF-based measures (such as entropy or the Kullback-Leibler divergence) without computing explicitly the PDFs. Although the kNN framework was mentioned in a semi-

nal work on Mean-Shift a while ago [24], it has rarely been used in image processing so far, except for high-dimensional clustering [25]. To our knowledge, no tracking method has never been proposed in this framework yet.

In term of comparison with other approaches, the proposed method can be characterized by such keywords as statistical, non-parametric, variable kernel bandwidth (kNN), joint color and geometry processing, and soft geometric constraint. *(i)* SAD, or similar non-robust and robust similarity measures, is deterministic in essence although it corresponds to solving the tracking problem with a parametric assumption on the residual PDF. The strict geometrical constraint does not allow much tolerance regarding motion model mismatch and the parametric PDF assumption prevents data fitting. *(ii)* kNN-KL can adapt to the data thanks to its non-parametric nature and the use of a variable kernel bandwidth. Because of its statistical point of view, it can account for some color variability of the ROI. Unfortunately, as it is well known, the absence of geometric constraint is a serious penalty. *(iii)* Pz-KL-G does include a soft geometrical constraint. However, the approximation of a PDF-based measure using a fixed kernel bandwidth, *i.e.*, without adjustment to the local density of the samples, is a weakness, as is clear from the experimental results. *(iv)* The Mean-Shift-based tracker used in the comparisons [16, 15] rely on another statistical measure: the Bhattacharya measure. Whether the differences observed between this tracker and the proposed method in the experimental results presented here depends on the measure itself or on the way geometry is involved[6] is unclear. Finally, *(v)* to a certain extent, the proposed method seems to provide answers to the problems *(i)* to *(iii)*.

Current works will focus on extending the kNN framework to other statistical measures such as mutual information [43] or the Bregman divergence [4]. Moreover, a study of the kernel bandwidth estimation approach, including the balloon estimation (kNN), the sample point estimation [17], or a hybrid estimation [39], will be carried out.

A Parzen windowing method and limitations

The duality between (7) (number of samples in a fixed volume) and (8) (volume necessary to contain a fixed number of samples) appears clearly. These two approaches were compared in a simple situation: let P and Q be two sets of samples of $\mathbb{R}^d$ drawn from two normal laws of mean 128 in each dimension and 132 in each dimension, resp., and two different, randomly generated, diagonally dominant covariance matrices. The Kullback-Leibler divergence $\mathcal{D}_{KL}(P,Q)$ was estimated using the Parzen-based method of toolbox [29] and using the kNN-based expression (19) choosing $k = 3$. Two series of experiments were performed: *(i)* with a fixed sample size $|P| = |Q| = 1000$ and the feature space dimension varying between 1 and 10; and *(ii)* with a fixed dimension $d = 5$ and a sample size[7] varying between 100 and 10000. The estimations were repeated 100 times (*i.e.*, using 100 different sample drawings for each set of parameters) and averaged (see Fig. 15)[8]. These experiments illustrate the curse of dimensionality and suggest that the kNN framework is better adapted to estimate the Kullback-Leibler divergence than the Parzen approach, especially in high dimension and even with few samples.

B Derivative of the Kullback-Leibler divergence

B.1 Expression

The Kullback-Leibler divergence is equal to

$$\mathfrak{D}_{\mathrm{KL}}(f_{T_\varphi}, f_R) = H^{\times}(f_{T_\varphi}, f_R) - H(f_{T_\varphi}) \tag{26}$$

where the cross entropy $H^{\times}(f_{T_\varphi}, f_R)$ can be approximated by (13)

$$H^{\times}(f_{T_\varphi}, f_R) \simeq -\frac{1}{|T_\varphi|} \sum_{s \in T_\varphi} \log f_R(s), \tag{27}$$

and the differential entropy $H(f_{T_\varphi})$ can be approximated by the Ahmad-Lin estimator [1]

$$H_{\mathrm{AL}}(T_\varphi) = -\frac{1}{|T_\varphi|} \sum_{s \in T_\varphi} \log f_{T_\varphi}(s). \tag{28}$$

In (28), the PDF is by definition equal to

$$f_{T_\varphi}(s) = \frac{1}{|T_\varphi|} \sum_{t \in T_\varphi} K_\sigma(s - t). \tag{29}$$

The same estimation (replacing T_φ with R) will be used in (27).

Therefore, we have

$$\mathcal{E}(\varphi) := \sum_{s \in T_\varphi} \log f_R(s) - \log f_{T_\varphi}(s) \tag{30}$$

$$\simeq -|T_\varphi| \; \mathfrak{D}_{\mathrm{KL}}(f_{T_\varphi}, f_R). \tag{31}$$

[6] A Gaussian weighting of the features according to their distance to the center of the ROI (which can be seen as a radial layout constraint) for the Mean-Shift-based tracker versus a joint radiometric/geometric processing for kNN-KL-G.

[7] Still with the condition $|P| = |Q|$.

[8] Note that the covariance matrices in Experiment *(i)* (*i.e.*, $|P| = |Q| = 1000$) for $d = 5$ and Experiment *(ii)* (*i.e.*, $d = 5$) for $|P| = |Q| = 1000$ were different. Therefore, the corresponding divergences do not match.

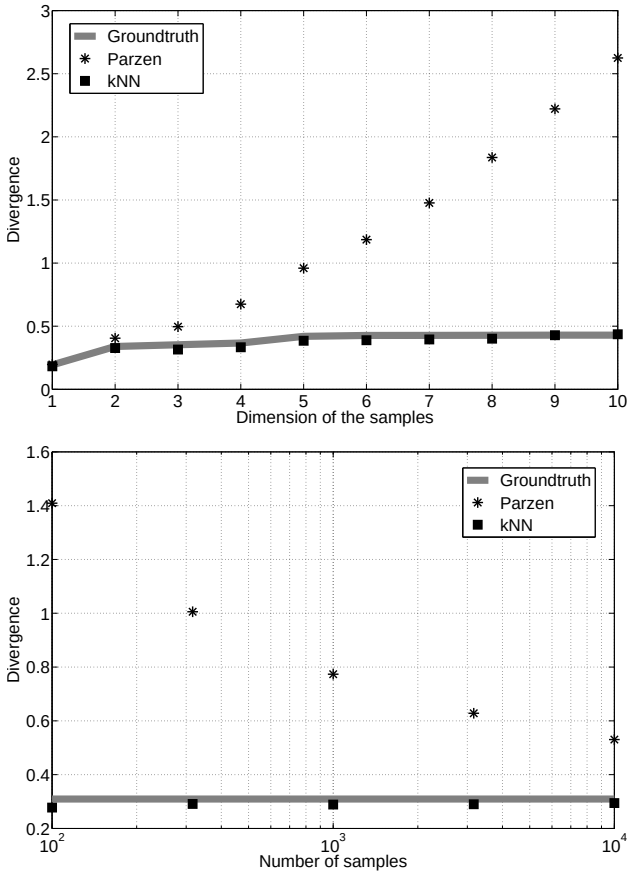


Figure 15: Comparison of the estimations of the Kullback-Leibler divergence between two normal laws using a Parzen-based method and the kNN approach. Theoretical divergence: gray line; Parzen-based estimation: star-shaped dots; kNN-based estimation: square-shaped dots. *(in lexicographical order)* Fixed sample size (1000) and varying dimension; Fixed dimension (5) and varying sample size.

Note that $|T_\varphi|$ is constant for all candidate regions in a given frame. Consequently, taking the derivative of (30) with respect to φ does not require to care about the interval of summation. Let the transformation φ be a translation (u, v) combined with a scaling by α. Sample set T_φ is equal to

$$T_\varphi = \{(I_{\text{tgt}}(x+u, y+v), x/\alpha, y/\alpha),\ (x,y) \in \Omega\}. \quad (32)$$

The derivative of (30) with respect to $\varphi = (\alpha, u, v)$ is equal to

$$\nabla\mathcal{E}(\varphi) = \sum_{s\in T_\varphi} \left(\frac{1}{f_R(s)} \frac{1}{|R|} \sum_{t\in R} \frac{\partial}{\partial\varphi} K_\sigma(s-t) - \frac{1}{f_{T_\varphi}(s)} \frac{1}{|T_\varphi|} \sum_{t\in T_\varphi} \frac{\partial}{\partial\varphi} K_\sigma(s-t) \right) \quad (33)$$

$$= \sum_{s\in T_\varphi} \left(\frac{1}{f_R(s)} \frac{1}{|R|} \sum_{t\in R} \mathcal{D}_s(T_\varphi)\ \nabla K_\sigma(s-t) - \frac{1}{f_{T_\varphi}(s)} \frac{1}{|T_\varphi|} \sum_{t\in T_\varphi} \frac{\partial}{\partial\varphi} K_\sigma(s-t) \right) \quad (34)$$

where

$$\mathcal{D}_s(T_\varphi) = \begin{bmatrix} 0 & 0 \\ \nabla I^Y_{\text{tgt}} \begin{pmatrix} s_x+u \\ s_y+v \end{pmatrix} & \nabla I^U_{\text{tgt}} \begin{pmatrix} s_x+u \\ s_y+v \end{pmatrix} \\ 0 & -\frac{1}{\alpha^2}[s_x \quad s_y] \\ \nabla I^V_{\text{tgt}} \begin{pmatrix} s_x+u \\ s_y+v \end{pmatrix} & \left[\ 0\ \right]_{[2\times 2]} \end{bmatrix} \quad (35)$$

Matrix $\mathcal{D}_s$ has p lines corresponding to the number of parameters of the motion model φ and d columns corresponding to the dimension of the feature space (here, (Y, U, V, x, y)). After some steps, one gets

$$\nabla\mathcal{E}(\varphi) = \sum_{s\in T_\varphi} \mathcal{D}_s(T_\varphi) \left(\frac{\nabla f_R(s)}{f_R(s)} - \frac{\nabla f_{T_\varphi}(s)}{f_{T_\varphi}(s)} + \frac{1}{|T_\varphi|} \sum_{t\in T_\varphi} \frac{\nabla K_\sigma(t-s)}{f_{T_\varphi}(t)} \right). \quad (36)$$

B.2 Term interpretation

Let us focus on the following term of (36)

$$\mathcal{A}(s) := \frac{1}{|T_\varphi|} \sum_{t\in T_\varphi} \frac{\nabla K_\sigma(t-s)}{f_{T_\varphi}(t)}. \quad (37)$$

When the number of samples $|T_\varphi|$ tends toward infinity, $\mathcal{A}$ tends toward

$$\mathcal{A}_\infty(s) = \int_{\mathbb{R}^d} f_{T_\varphi}(t)\ \frac{\nabla K_\sigma(t-s)}{f_{T_\varphi}(t)}\ \mathrm{d}t. \quad (38)$$

Kernel K_σ is radially symmetric. Therefore, for all x and y such that $x = -y$, we have

$$\nabla K_\sigma(x) = -\nabla K_\sigma(y). \quad (39)$$

Consequently, (38) convergences (at least weakly) toward zero.

C Derivative of the divergence: Mean-Shift-based approximation and kNN implementation

C.1 Mean-Shift

In the context of tracking, Mean-Shift is often used to refer to a Mean-Shift-based algorithm. Here, it refers to the original meaning of approximation of $\nabla f/f$ using the shift from the mean of neighboring samples [24, 23][9]

$$\frac{\nabla f(s)}{f(s)} \simeq \frac{d+2}{\sigma^2}(\bar{s}_\sigma - s) \tag{40}$$

where

$$\bar{s}_\sigma = \frac{1}{n} \sum_{t \in W_\sigma(s)} t \tag{41}$$

is the mean of the samples (which happen to be n in number) contained in a window W_σ of radius σ centered at s. If f is a normal distribution with mean μ and variance σ^2, then the Mean-Shift has the following, simple analytical expression

$$\frac{\nabla f(s)}{f(s)} = \frac{\mu - s}{\sigma^2}. \tag{42}$$

C.2 kNN-based expression

The first two terms enclosed in parentheses in (36) can be approximated using the Mean-Shift (40). The expression of the mean (41) can be replaced with its kNN equivalent [24]

$$\bar{s}_{\rho_k(s)} = \frac{1}{k} \sum_{t \in W_{\rho_k(s)}} t. \tag{43}$$

In the third term enclosed in parentheses in (36), the PDF f_{T_φ} can also be replaced with its kNN expression (8). Therefore, using the Mean-Shift approximation, the derivative of the Kullback-Leibler divergence can be written as a kNN-based expression

$$\begin{aligned} k\ \nabla\mathcal{E}(\varphi) = \sum_{s \in T_\varphi} \mathcal{D}_s(T_\varphi) \Bigg(& \frac{d+2}{\rho_k^2(R,s)} \sum_{t \in W_{\rho_k(R,s)}} (t-s) \\ & - \frac{d+2}{\rho_k^2(T_\varphi,s)} \sum_{t \in W_{\rho_k(T_\varphi,s)}} (t-s) \\ & + v_d \sum_{t \in T_\varphi} \rho_k^d(T_\varphi,t)\ \nabla K_{\rho_k(T_\varphi,t)}(t-s) \Bigg) \end{aligned} \tag{44}$$

[9] In [23], see page 534.

where $K_{\rho_k(T_\varphi,t)}(\cdot - s)$ is a window of radius $\rho_k(T_\varphi,t)$ centered at s.

C.3 Term approximation

Let us now focus on the following term of (44) (which corresponds to the kNN version of (37))

$$\mathcal{A}_{\text{kNN}}(s) := \sum_{t \in T_\varphi} \rho_k^d(T_\varphi,t)\ \nabla K_{\rho_k(T_\varphi,t)}(t-s). \tag{45}$$

In light of Appendix B.2, this term could be neglected if $|T_\varphi|$ is large enough. Nevertheless, let us propose an approximation of it.

Window $K_{\rho_k(T_\varphi,t)}(\cdot - s)$ at t is equal to $1/(\rho_k^d(T_\varphi,t)\ v_d)$ if $|t-s| \leq \rho_k(T_\varphi,t)$ and zero otherwise. A finite difference approximation can be used to write

$$\nabla K_{\rho_k(T_\varphi,t)}(t-s) = \begin{cases} \frac{1}{\rho_k^d(T_\varphi,t)\ v_d}\ \frac{s-t}{|s-t|} & \text{if } |s-t| = \rho_k(T_\varphi,t) \\ 0 & \text{otherwise.} \end{cases} \tag{46}$$

Therefore, term (45) can be approximated by

$$\mathcal{A}_{\text{kNN}}(s) \simeq \frac{1}{v_d} \sum_{\substack{t \in T_\varphi \\ |t-s| = \rho_k(T_\varphi,t)}} \frac{s-t}{\rho_k(T_\varphi,t)}. \tag{47}$$

This approximation leads to the final expression (25) of the kNN-based derivative of (19). Note that, in practice, the summation condition $|t-s| = \rho_k(T_\varphi,t)$ should be understood as $|t-s| \in \rho_k(T_\varphi,t) \pm \epsilon$ for a small ϵ.

References

[1] I. A. Ahmad and P. E. Lin. A nonparametric estimation of the entropy for absolutely continuous distributions. *IEEE Trans. Inform. Theory*, 22(3):372–375, 1976.

[2] G. Aubert, M. Barlaud, O. Faugeras, and S. Jehan-Besson. Image segmentation using active contours: Calculus of variations or shape gradients? *SIAM J. Appl. Math.*, 63(6):2128–2154, 2003.

[3] R. Venkatesh Babu, P. Pérez, and P. Bouthemy. Robust tracking with motion estimation and local kernel-based color modeling. *Image Vis. Comput.*, 25(8):1205–1216, 2007.

[4] A. Banerjee, S. Merugu, I. Dhillon, and J. Ghosh. Clustering with bregman divergences. *J. Mach. Learn. Res.*, 6:1705–1749, October 2005.

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

[6] S. Boltz, É. Debreuve, and M. Barlaud. High-dimensional statistical distance for region-of-interest tracking: Application to combining a soft geometric constraint with radiometry. In *International Conference on Computer Vision and Pattern Recognition*, Minneapolis (MN), USA, 2007.

[7] S. Boltz, A. Herbulot, É. Debreuve, M. Barlaud, and G. Aubert. Motion and appearance nonparametric joint entropy for video segmentation. *Int. J. Comput. Vision*, 80(2):242–259, 2008.

[8] S. Boltz, É. Wolsztynski, É. Debreuve, É. Thierry, M. Barlaud, and L. Pronzato. A minimum-entropy procedure for robust motion estimation. In *International Conference on Image Processing*, Atlanta (GA), USA, 2006.

[9] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In *European Conference on Computer Vision*, Prague, Czech Republic, 2004.

[10] T. Brox, M. Rousson, R. Deriche, and J. Weickert. Unsupervised segmentation incorporating colour, texture, and motion. In *Computer Analysis of Images and Patterns*, Groningen, The Netherlands, 2003.

[11] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel. A non-local algorithm for image denoising. In *International Conference on Computer Vision and Pattern Recognition*, San Diego (CA), USA, 2005.

[12] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel. A review of image denoising algorithms, with a new one. *SIAM Multiscale Model. Simul.*, 4(2):490–530, 2005.

[13] A. Bugeau and P. Pérez. Detection and segmentation of moving objects in highly dynamic scenes. In *International Conference on Computer Vision and Pattern Recognition*, Minneapolis (MN), USA, 2007.

[14] G. Carlsson, T. Ishkhanov, V. de Silva, and A. Zomorodian. On the local behavior of spaces of natural images. *Int. J. Comput. Vision*, 76(1):1–12, 2008.

[15] R. Collins, X. Zhou, and S. K. Teh. An open source tracking testbed and evaluation web site. In *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*, Breckenridge (CO), USA, 2005. Code available at: http://www.vividevaluation.ri.cmu.edu/.

[16] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *International Conference on Computer Vision and Pattern Recognition*, Hilton Head Island (SC), USA, 2000.

[17] Dorin Comaniciu. An algorithm for data-driven bandwidth selection. *IEEE Trans. Pattern Anal.*, 25(2):281–288, 2003.

[18] J. Costa and A. O. Hero. Manifold learning using euclidean K-nearest neighbor graphs. In *International Conference on Acoustics, Speech, and Signal Processing*, Montreal, Canada, 2004.

[19] D. Cremers, M. Rousson, and R. Deriche. A review of statistical approaches to level set segmentation: integrating color, texture, motion and shape. *Int. J. Comput. Vision*, 72(2):195–215, 2007.

[20] A. Elgammal, R. Duraiswami, and L. S. Davis. Probabilistic tracking in joint feature-spatial spaces. In *International Conference on Computer Vision and Pattern Recognition*, Madison (WI), USA, 2003.

[21] Excerpt from the movie "Poltergay" directed by Eric Lavaine. Produced by Fabio Conversi, François Cornuau, and Vincent Roget, 2006.

[22] E. Fix and J. L. Hodges. Discriminatory analysis, non-parametric discrimination: consistency properties. Technical Report 4, USAF School of Aviation Medicine, Randolph Field, 1951.

[23] K. Fukunaga. *Introduction to statistical pattern recognition (2nd Ed.)*. Academic Press Professional, Inc., 1990.

[24] K. Fukunaga and L. D. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Trans. Inform. Theory*, 21(1):32–40, 1975.

[25] Bogdan Georgescu, Ilan Shimshoni, and Peter Meer. Mean shift based clustering in high dimensions: A texture classification example. In *International Conference on Computer Vision*, Nice, France, 2003.

[26] M. N. Goria, N. N. Leonenko, V. V. Mergel, and P. L. Novi Inverardi. A new class of random vector entropy estimators and its applications in testing statistical hypotheses. *J. Nonparametr. Stat.*, 17(3):277–297, 2005.

[27] DPI Göttingen, TSTOOL toolbox for nearest neighbor statistics, 1997. Code available at: http://www.dpi.physik.uni-goettingen.de/tstool/.

[28] A. Herbulot, S. Jehan-Besson, S. Duffner, M. Barlaud, and G. Aubert. Segmentation of vectorial image features using shape gradients and information measures. *J. Math. Imaging Vis.*, 25(3):365–386, 2006.

[29] Alexander Ihler. Kernel density estimation toolbox for Matlab, 2003. Code available at: http://ttic.uchicago.edu/~ihler/code/kde.php.

[30] J. Kim, J. W. F. Fisher, A. Yezzi, M. Çetin, and A. S. Willsky. Nonparametric methods for image segmentation using information theory and curve evolution. *IEEE Trans. Image Process.*, 14(10):1486–1502, 2005.

[31] L. Kozachenko and N. Leonenko. On statistical estimation of entropy of random vector. *Problems Infor. Transmiss.*, 23(2):95–101, 1987. (translated from Problemy Peredachi Informatsii, in Russian, vol. 23, No. 2, pp. 9-16, 1987).

[32] Ann B. Lee, Kim S. Pedersen, and David Mumford. The nonlinear statistics of high-contrast patches in natural images. *Int. J. Comput. Vision*, 54(1-3):83–103, 2003.

[33] J. Lin. Divergence measures based on the shannon entropy. *IEEE Trans. Inform. Theory*, 37(1):145–151, 1991.

[34] D. O. Loftsgaarden and C. P. Quesenberry. A nonparametric estimate of a multivariate density function. *Ann. Math. Statist.*, 36(3):1049–1051, 1965.

[35] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, 2004.

[36] Tom Minka. Divergence measures and message passing. Technical Report MSR-TR-2005-17, Microsoft Research, 2005.

[37] N. Paragios and R. Deriche. Geodesic active regions and level set methods for supervised texture segmentation. *Int. J. Comput. Vision*, 46(3):223–247, 2002.

[38] P. Pérez, C. Hue, J. Vermaak, and M. Gangnet. Color-based probabilistic tracking. In *European Conference on Computer Vision*, Copenhaguen, Denmark, 2002.

[39] Stephan R. Sain. Multivariate locally adaptive density estimation. *Comput. Stat. Data Anal.*, 39(2):165–186, 2002.

[40] D. W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley, 1992.

[41] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London, 1986.

[42] G. R. Terrell and D. W. Scott. Variable kernel density estimation. *Ann. Statist.*, 20(3):1236–1265, 1992.

[43] P. Viola and W. M. Wells. Alignment by maximization of mutual information. *Int. J. Comput. Vision*, 24(2):137–154, 1997.

[44] J. Weickert and C. Schnörr. Variational optic flow computation with a spatio-temporal smoothness constraint. *J. Math. Imaging Vis.*, 14(3):245–255, 2001.

[45] Changjiang Yang, Ramani Duraiswami, Nail A. Gumerov, and Larry Davis. Improved fast gauss transform and efficient kernel density estimation. In *International Conference on Computer Vision*, Nice, France, 2003.

[46] Shan Zhu and Kai-Kuang Ma. A new diamond search algorithm for fast block-matching motion estimation. *IEEE Trans. Image Process.*, 9(2):287–290, 2000.