PRISM: PRincipled Implicit Shape Model

Alain Lehmann[1]
lehmann@vision.ee.ethz.ch
Bastian Leibe[2]
leibe@umic.rwth-aachen.de
Luc van Gool[1]
vangool@vision.ee.ethz.ch

[1] Computer Vision Laboratory
ETH Zurich
Switzerland
[2] Mobile Multimedia Processing
UMIC Research Centre
RWTH Aachen University

Most current state-of-the-art detectors are based on either the sliding-window (*e.g.* [1]) or the Generalised Hough transform paradigm (*e.g.* [2]). This paper bridges the gap between the two paradigms and leverages their advantages. Our framework benefits from the sound sliding-window reasoning and provides a well-grounded mathematical explanation to the voting procedure of the Implicit Shape Model (ISM) [2]. It thereby overcomes the questionable *marginalisation over features* justification. Moreover, it allows for discriminative voting weights which was not possible in ISM. We present a Gaussian Mixture Model implementation of our framework which achieves state-of-the-art performance.

The second key contribution deals with commonly used heuristics such as soft-matching and spatial pyramid descriptors. Both can be expressed formally within our framework. More importantly, we show that they can be avoided during detection without loosing their positive effect. This is achieved by moving them entirely to the learning stage where they act as model regularisation.

Sliding-Windows meet the Hough Transform.

Object detection by means of sliding-windows forms the basis of our framework. It formulates object detection as a search problem. Given a new image I, the goal is to find the best hypothesis

$$\lambda^* = \arg\max_{\lambda \in \Lambda} S(\phi(\lambda, I)|W) \tag{1}$$

from the search space Λ. ϕ is a windowing function which crops out sub-images and S is the score function which ranks these hypotheses according to a learned model W. Although this process is massively parallel, most implementations process one hypothesis after the other which motivates the term *sliding* window.

Footprint & Invariants. A key contribution of this work is to extend the notion of "windowing" to what we call the object footprint ϕ. The basic idea is to compute a representation of the object which is independent of its position and size (and pose, *etc.*) in the image. This is achieved by computing invariants $\mathbb{I}(\lambda, f)$ of extracted image-features f and the object-hypothesis λ. The use of such invariants makes the modelling of geometrical transformations explicit. This is an important aspect when it comes to *e.g.* view-invariant object recognition.

Hough-Transform. The final step towards the Hough transform is to use a linear model which scores hypotheses by (details omitted)

$$\langle \phi(\lambda, I), W \rangle = \sum_{f \in \mathscr{F}(I)} f_\omega \cdot W(f_c, \mathbb{I}(\lambda, f)), \tag{2}$$

where $\mathscr{F}(I)$ is the set of features (extracted from the image), f_c is the index of the best-matching visual word (in a learned vocabulary), and f_ω is a feature weight (*e.g.* accounting for matching quality). This form makes the connection to the Hough transform explicit. The function $W(f_c, \mathbb{I}(\cdot, f))$ acts as *voting pattern*. It is simply a re-parametrisation of the model weights.

Discussion & Benefits. Our score is structurally similar to the probabilistic formulation of ISM [2] , *i.e.*

$$p(\lambda|I) = \sum_{f \in \mathscr{F}(I)} p(f|I) \cdot p(\lambda|f), \tag{3}$$

where the summation is explained by means of marginalisation. However, as we will explain, this argument is not justified. Our framework avoids such questionable arguments by following a clean sliding-window reasoning. Compared to ISM's density-based formulation (which imposes a non-negativity and normalisation constraint), our model is unconstrained. Thus, the voting weights W can be *negative* and learned in a discriminative fashion. This is an important advantage in practice as most state-of-the-art detectors rely on discriminative learning.

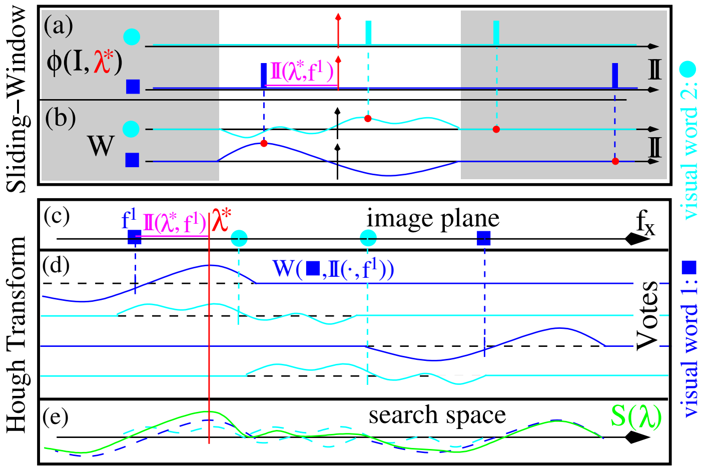

Figure 1: *PRISM: From the Sliding-Window (top) to the Hough Transform paradigm (bottom). 1D example with two visual words (blue rectangle, cyan circle). A footprint (a) gets extracted for a fixed hypothesis λ^* (red). The inner product with the weight function (b) results in a sum of point evaluations (red dots). Various features do* not *affect the score (shaded). Each extracted feature (c) casts a voting pattern (d) which, summed up, result in the final hypothesis score (e).*

Implementation. Our implementation is based on Gaussian Mixture Models for the weights W. The aim is to overcome ISM's strong dependence on the training data size due to its non-parametric kernel density estimators. In such a setting, a clever choice of the invariant $\mathbb{I}$ is important, which we will discuss. The search for hypotheses is implemented through gradient based search techniques which relates to ISM's mean-shift search. We further present a simple mixture-dropping heuristic which leads to a significant speed-up.

Fast Soft-Matching.

Soft-matching is a heuristic where a feature activates multiple codewords from the visual vocabulary, instead of just the best matching one. This has proven very effective to increase robustness against quantisation effects [3]. Another common practice is the use of spatial-histogram *pyramids* instead of flat histograms. Unfortunately, the downside of both is an increase of computational costs, due to more emitted votes or increased dimensionality (of the pyramid). We show that both can be avoided during detection without missing their positive effect. We argue that both heuristics lead to model regularisation which should be done entirely during learning. This then allows for simpler and faster detection systems.

The simple, yet effective trick is to consider a linear mapping $\phi \mapsto B\phi$ of the footprint. In case of the spatial pyramid, B is a matrix which constructs the pyramid from the flat histogram. Then, the score is computed as $\langle B\phi, W \rangle = \langle \phi, B'W \rangle$ where $B'W$ can be pre-computed during training. The left-hand side is the usual pyramid-matching while the right-hand side operates in the *low*-dimensional (flat) histogram space. For soft-matching, B can be interpreted as blurring of the single (per feature) Dirac of the footprint. This results in multiple codeword activation with decreasing weights. Following this interpretation, $B'W$ can be thought of as a blurring of the learned model.

We experimentally show that soft-matching during detection only degrades results if blurring during learning was sufficiently strong. Hence, fast nearest-neighbour matching can be used for detection without loss of performance.

[1] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[2] B. Leibe, A. Leonardis, and B. Schiele. Robust object detection by interleaving categorization and segmentation. *IJCV*, 77(1-3):259–289, 2008.

[3] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In CVPR 2008.