České vysoké učení technické v Praze
Fakulta elektrotechnická

Czech Technical University in Prague
Faculty of Electrical Engineering

Dr. Ing. Jiří Matas

Metoda lokálních afinních rámců pro robustní rozpoznávání velkých sad objektů

Scalable and Robust Object Recognition with the Local Affine Frame Method

Summary

Realistic approaches to large scale object recognition, i.e. for detection and localisation of hundreds or more objects, must support sub-linear time indexing. In the paper, we propose a method capable of recognising one of N objects in $log(N)$ time.

The "visual memory" is organised as a binary decision tree that is built to minimise average time to decision. Leaves of the tree represent a few local patches, and each non-terminal node is associated with a 'weak classifier'. In the recognition phase, a single invariant measurement on a query patch decides in which subtree a corresponding patch is sought.

The method possesses all the strengths of local affine region methods – robustness to background clutter, occlusion, and large changes of viewpoints. We show that it supports near real-time recognition of hundreds of objects with state-of-the-art recognition rates. After the test image is processed (in a second on a current PCs), the recognition via indexing into the visual memory requires milliseconds. making.

Souhrn

Metody, které si kladou za cíl vizuální rozpoznávání velkých souborů objektů, t.j. metody pro detekci a lokalizaci stovek a více objektů, se neobejdou bez indexace, která podporuje vyhledávání v sub-lineárním čase. V přednášce popisujeme metodu schopnou rozpoznat jeden z N objektů v čase úměrném $log(N)$.

"Vizuální paměť" je organizována jako binární rozhodovací strom, který je vystavěn tak, že se minimalizuje střední doba rozhodnutí. Listy stromu obsahují malý počet výřezů z obrázků. Každý neterminální uzel obsahuje rozhodovací pravidlo třídící výřezy jasově a geometricky normalizované metodou lokálních afinních rámců.

Navržená metoda rychlého vyhledávání objektů si zachovává všechny výhodné vlastnosti metod založených na afinních rámcích - robustnost vůči zákrytu, změně na pozadí, změně úhlu pohledu a změně osvětlení. Experimenty ukazují, že metoda dosahuje na standardních testovacích problémech výsledků, které jsou lepší, než nejlepší publikované. Na PC s 2GHz procesorem trvá zpracování vstupního obrázku velikosti 640x480 asi jednu sekundu. Čas indexace je zanedbatelný, vyžaduje desítky milisekund.

Klíčová slova: rozpoznávání objektů, vizuální rozpoznávání, registrace obrázků, lokální afinní rámce, maximálně stabilní extremální oblasti

Keywords: object recognition, visual recognition, image registration, local affine frames, maximally stable extremal regions

Contents

1 Introduction

In recent years, research in object recognition has progressed rapidly. Methods based on correspondences of invariantly detected regions have achieved robustness to background clutter, occlusion, and large changes of viewpoint. Impressive results, albeit for certain classes of objects, have been reported [11, 3, 5].

Realistic approaches to recognition, detection and localisation of objects from large collections must support sub-linear indexing, i.e. the ability to associate current visual input with objects represented in the memory, at a speed that does not significantly depend on the number of images and objects already seen. Any technique that compares the current visual input one-by-one with stored models is linear in the number of known objects. Such recognition techniques, solving effectively a sequence of two-image matching problems, will have, sooner or later, an unacceptable response time. Searching and indexing are well-studied subjects, and two sub-linear methods dominate the field – hashing and tree search. This paper presents an approach that achieves sub-linear, real-time recall by representing the visual memory as a binary decision tree organised to minimise average time to decision.

The novel features of the proposed method, and the method itself, is easier explained if the reader is familiar with the state-of-the-art approach of D. Lowe [5]. In Lowe's recognition method, processing of an image starts by extracting square patches invariant to similarity transformations. Next, each patch is described by the SIFT feature – a 128-dimensional vector consisting of sixteen eight-bin weighted histograms of gradient orientations. In the training stage, SIFT descriptors from all training images are organised in a kD-tree. Correspondence between patches from the training images and a query image are established as follows. First, SIFT descriptors are computed on the query image. For each SIFT, the kD-tree returns one stored patch if the tree contains a descriptor that is significantly closer to the queried descriptor than other stored descriptors, else no match is reported. The matched pairs of the query and kD-tree patches form tentative correspondences, which are confirmed or rejected in subsequent verification and consistency checks (these are not relevant for this paper). Lowe's method can be summarised as: (i) detect local coordinate frames in an invariant manner, (ii) represent them by a fix-sized feature vectors, (iii) search efficiently for nearest neighbours of the vectors and (iv) find geometrically consistent groups of correspondences of local coordinate frames.

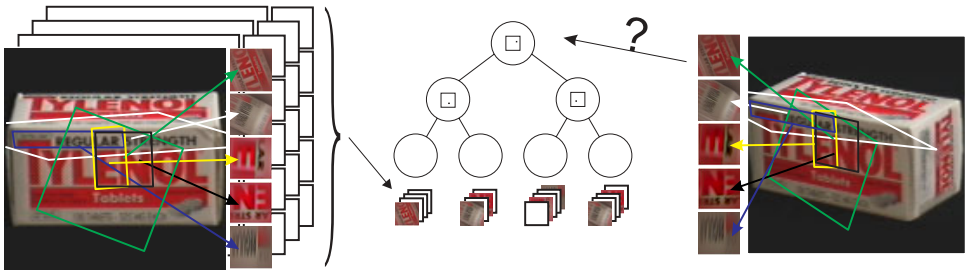


Figure 1: An example of features stored in the visual memory.

The first step of the proposed LAF-TREE method is in principle the same as in Lowe's approach. We chose a different type of distinguished regions (transformation covariant regions) – the maximally stable extremal regions (MSERs, [6]), but any affine (scale) invariant processes could be used[1]. A formal definition of a distinguisehd region is:

[1]Executables of a number of covariant region detectors (including MSERs) are available on the web

Definition 1 Distinguished region. *Let image I be a mapping $I : \mathcal{D} \subset \mathbb{Z}^2 \rightarrow \mathcal{S}$. Let $\mathcal{P} \subset 2^{\mathcal{D}}$, i.e. $\mathcal{P}$ is a subset of the power set (set of all subsets) of $\mathcal{D}$. Let $\mathcal{A} \subset \mathcal{P} \times \mathcal{P}$ be an adjacency relation on $\mathcal{P}$ and let $f : \mathcal{P} \rightarrow \mathcal{T}$ be any function defined on $\mathcal{P}$ with a totally ordered range $\mathcal{T}$. A region $\mathcal{Q} \in \mathcal{P}$ is distinguished with respect to function f iff $f(\mathcal{Q}) > f(\mathcal{Q}'), \forall(\mathcal{Q}, \mathcal{Q}') \in \mathcal{A}$.*

The LAF-TREE method establishes local affine frames (LAFs) by constructions described in [8]. The structure of the method is visualised in Fig. **??** and summarised in Algorithm 1:

Algorithm 1: Structure of the proposed MSER-LAF method

---

1. For every database and query image, compute affine-covariant regions of data-dependent shape.
2. Construct local affine frames (LAFs) on the regions using several affine-covariant constructions.
3. Generate intensity representations of local image patches normalised according to the local affine frames. Photometrically normalise the patches.
4. Establish tentative correspondences between frames of query and database images. Compute similarity between the patches, select most similar pairs.
5. Find a globally consistent subset of the correspondences. Infer the presence and location of the objects.

---

but this is a superficial difference - different detectors and frame constructions can be easily combined or replaced. The novelty of the LAF-TREE approach is in the departure, in Step (ii), from the "compute a fixed-size feature vector on a fixed measurement region" paradigm. In this paradigm, the local reference frame is described by a function of pixel values from a measurement region whose size and shape, if expressed in the local frame coordinates, is fixed. In Lowe's approach, the shape is a square of a predefined size. It is clear that a fixed measurement region will lead to difficulties when recognising certain classes of objects, e.g. "wire-like" objects as bicycles where any square neighbourhood includes background. Perhaps more significantly, a measurement region of a certain size will be too big for some frames, e.g. including parts of background or discontinuities, and yet it will be too small for other frames whose descriptors will not be discriminative.

The problem of a better-than-fixed measurement region seems insurmountable. How can possibly be the measurement region adapted unless we know what we are looking at? *We finesse the problem by interleaving the processes of recognising the frame and deciding where to measure next.* The frame is recognised by descending a decision-measurement tree where each decision not only reduces the number of potential corresponding frames represented in the tree but also defines which measurements are taken next. More precisely, a binary tree is formed in the learning stage. For each non-terminal node, a binary valued measurement-decision function, called a 'weak classifier', is selected from a large pool according to an optimisation criterion. The criterion is a lower bound on the expected

at `http://www.robots.ox.ac.uk/~vgg/research/affine`

time to decision. The term 'weak classifier' stresses the obvious analogies with discrete AdaBoost - it is selected by a greedy algorithm, it could be any binary function of pixel values and, as will be shown later, it is not required to make unequivocal decisions.

Establishing tentative correspondences with the decision-measurement tree has a number of favourable properties. The advantages of a data-specific measurement region have already been mentioned. From a computational point of view, efficiency of recognition is increased since only a small fraction of potential measurements is evaluated. In case that a measurement is close to a decision boundary, or not available at all as in the case when it is taken from the background, robustness of the search is easily achieved by inserting the training frame in both subtrees. With this modification, the search in the recognition stage descend always into only a single branch, guaranteeing that a leaf of the tree is reached in $\log(N)$ steps, where N is the number of frame instances stored in the tree. Last but not least, the learning process explicitly takes into account geometric uncertainty and image statistics to minimise the response time (see Section 2). The final recognition Step (iv), verification of the presence of objects by finding geometrically consistent groups of correspondences, is not time-critical, since the number of tentative matches per frame is small. It can be implemented by RANSAC or a voting scheme.

1.1 Related work.

Our work on decision trees was inspired by Lepetit and Fua [3], who were the first to introduce decision trees for the recall of local image representation. Their approach, however, differs from ours in several areas. First, they set the tree size (and the number of trees, since they are using multiple randomised trees) by hand, while in our approach the tree size is a function of image database content. Next, our measurements are invariant to affine deformations of the image (thanks to the LAF constructions), we thus do not need a 3D model or synthetically warped 2D images to capture the appearance variations. We also explicitly consider image noise and background segmentation of the measurements, while Lepetit et al. synthetically generate noisy patches and patches with random background. We present experiments on datasets containing hundreds of objects, while the results of Lepetit's (and also of Lowe's) work are demonstrated on only a few objects. The Video Google system by Sivic and Zisserman [11] is able of indexing of a full-length movie. A clustering of descriptors of local features is employed to reduce the recall time complexity by a constant factor. But the processing time is not really an issue, since everything is precomputed off-line, and the system is closed, i.e. the object queries must originate from images from the movie. The work by Nene and Nayar [7] supports real-time recognition using a space slicing search, but it is restricted to segmented objects. Neither object occlusion, cluttered background nor multiple objects in scene are supported.

The rest of the paper is structured as follows. Section 2 details the proposed decision-measurement tree structure. In Section 3, we experimentally show that the method supports near real-time recognition of hundreds of objects with state-of-the-art recognition rates. Section 4 concludes the paper.

2 Recognition with Decision-Measurement Trees

This section describes the decision-measurement tree which is used to represent the "visual memory". Generally, a decision tree is a tree structure where a simple test (a weak classifier) that splits the observation space is assigned to each non-terminal node. Each

leaf corresponds to a volume that is defined by the sequence of decisions made on the path from the tree root to that particular leaf. During the recall phase, the tree is traversed according to the decisions at non-terminal nodes, until a leaf node (and a corresponding volume) is reached. The elements in the reached volume do not necessarily match the query – being in the same volume does not imply proximity – and an additional evaluation of a similarity measure is necessary to distinguish matching and non-matching elements. Recall can be viewed as a sequential reduction of the set of candidate correspondences until a subset of a small predefined cardinality (called 'leaf capacity') is reached. The elements remaining in the subset are sequentially searched for matches.

Although we currently employ only one simple type of classifiers, multiple types can be freely combined within the tree. Our classifiers are binary functions $d_{\overline{\mathbf{x}},\Theta_{\overline{\mathbf{x}}}} : A \rightarrow \{\mathrm{L}, \mathrm{R}\}$, which threshold a single pixel value. $\overline{\mathbf{x}}$ is a vector specifying the measurement location, $\Theta_{\overline{\mathbf{x}}}$ is a scalar threshold on value at $\overline{\mathbf{x}}$, A is a local affine frame, and $\{\mathrm{L}, \mathrm{R}\}$ are the decisions to search Left and Right subtrees respectively.

Due to image noise, the decisions are ambiguous for values close to the thresholds $\Theta_{\overline{\mathbf{x}}}$. The ambiguity can be solved in the recognition phase by descending both subtrees, as e.g. in the classical kD-tree algorithm. In an alternative approach, the elements are in ambiguous cases stored redundantly in both subtrees. There is then no need to backtrack or split the tree search during recognition (recall), all uncertainties are solved in the training phase. This approach allows for faster retrieval at the expense of memory needed for the redundant representation. Since our motivation is to achieve high recognition speeds, we have adopted the second approach. The design leads to a straightforward retrieval algorithm (see Algorithm 1). The retrieval is very fast since for each query frame A only one evaluation of a weak classifier (thresholding of a single pixel value) is performed at each tree level. The depth of the tree is typically 15 to 25, depending on the database size.

2.1 Learning of the tree

(Algorithm 2). A separate tree is constructed for every type of LAF construction. Starting with a set S_A of frames of a single type of construction, the set is recursively divided into subsets at non-terminal nodes. Non-terminal nodes are inserted until (a) the cardinality of the particular subset is below a predefined threshold, the 'leaf capacity', or (b) the frames in the subset are indistinguishable. If either (a) or (b) is satisfied, a leaf node is constructed. The condition (b) accommodates for the situation where there are multiple images of the same object in the database, or when the objects contain repetitive structures.

Let $r(A)$ denote a random realisation of frame A in a query image. r is a random function that encapsulates geometric and photometric misalignments between corresponding frames, as well as image noise, blur and other image distortions. Algorithm 2 ensures that A is represented in every leaf where the probability of a query realisation $r(A)$ falling to that leaf is above a threshold Θ_p; Θ_p is a parameter of the method. Next we describe the process of evaluation of the probability that a query realisation $r(A)$ of frame A will descend the left ($p(d_{\overline{\mathbf{x}},\Theta_{\overline{\mathbf{x}}}}(r(A)) = \mathrm{L})$), and right ($p(d_{\overline{\mathbf{x}},\Theta_{\overline{\mathbf{x}}}}(r(A)) = \mathrm{R})$) subtree respectively, given a classifier $d_{\overline{\mathbf{x}},\Theta_{\overline{\mathbf{x}}}}$ and the frame A.

2.2 Estimation of geometric misalignment

. The precision of the geometric alignment achieved by the MSER-LAF method was analysed. Local affine frames were constructed on several image pairs related by known

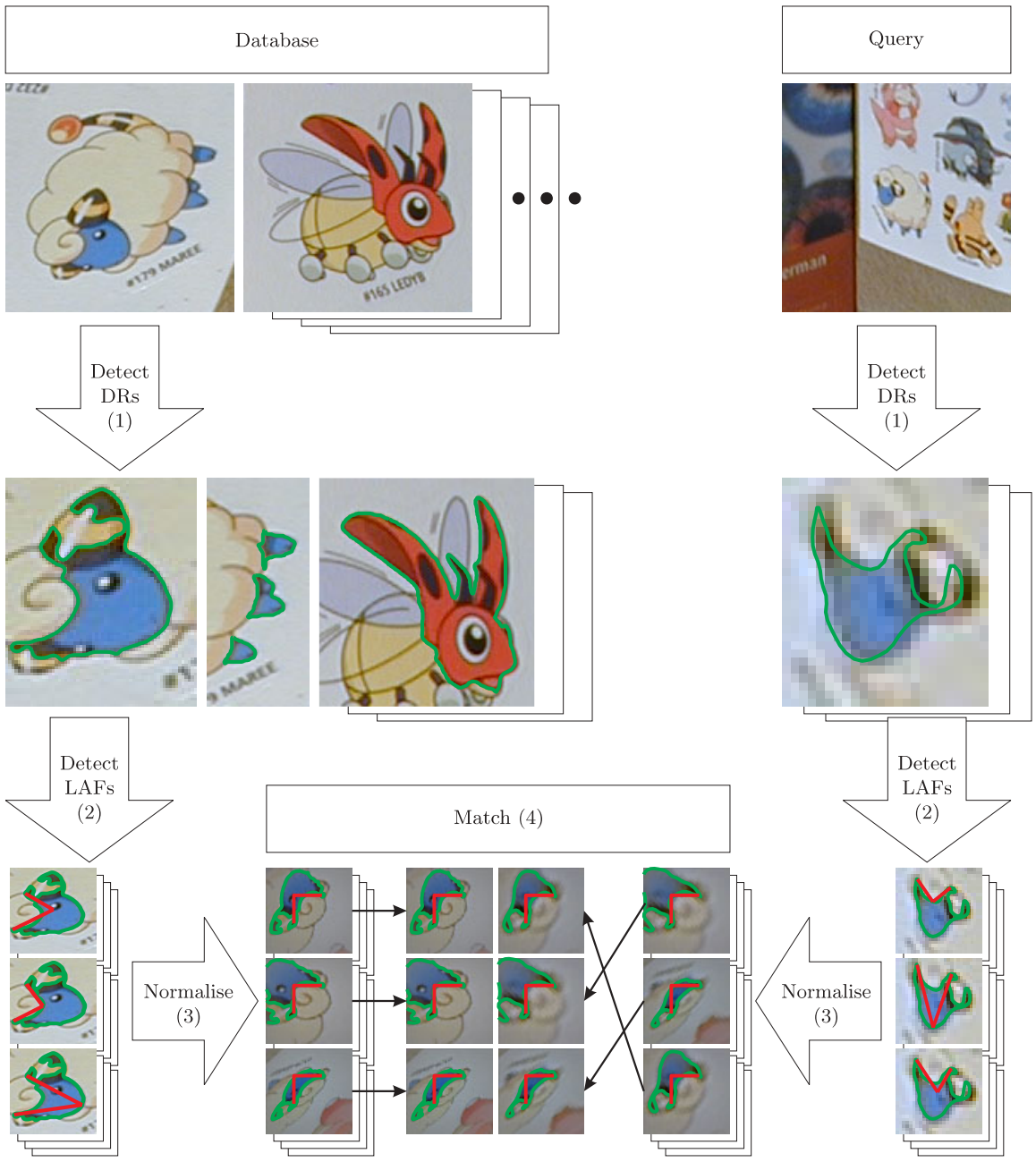


Figure 2: Structure of the MSER-LAF object recognition method

homographies. Corresponding frames did not align perfectly – a single spot in the scene occurs at slightly different pixels in the patches. Figure 3 shows covariance matrices of distributions of pixel displacements, estimated on thousands of frames[2]. The distributions represent a localisation uncertainty $l_{\overline{\mathbf{x}}}$ of pixels in query patches. As expected, the farther from the detected frame, the larger is the uncertainty. It is also seen that the distributions differ significantly for different types of frame constructions. A separate set of distributions is therefore maintained for each frame type.

[2]The patch size is 31×31 pixels, but for presentation clarity Figure 3 shows the distributions for only every second pixel of the patch

Algorithm 1: MSER-LAF-TREE: Retrieving stored frames

Input:
A: a query local affine frame
Output:
S: a set of candidate matches

Tree.retrieveFrames $(A) \rightarrow S$
$S :=$ root.retrieveFrames (A)

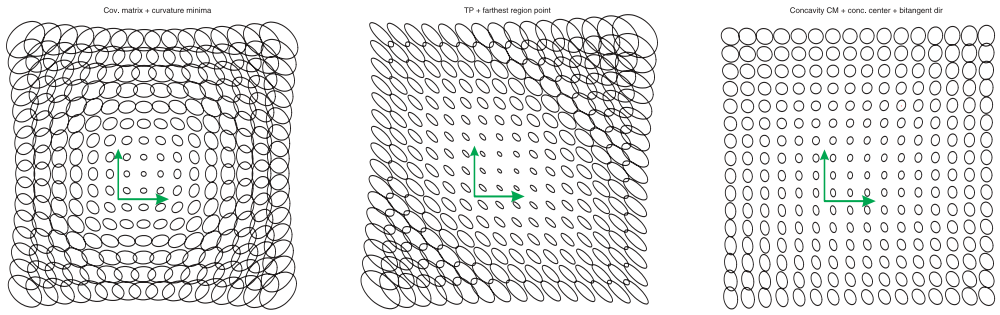
Node.retrieveFrames $(A) \rightarrow S$
if isLeaf **then**
$S := \{A_i : A_i \in \text{leafFrames} \wedge \text{similar}(A, A_i)\}$
else
if $d_{\overline{\mathbf{x}},\Theta_{\overline{\mathbf{x}}}}(A) = \text{L}$ **then**
$S :=$ leftSubtree.retrieveFrames (A)
else $\{d_{\overline{\mathbf{x}},\Theta_{\overline{\mathbf{x}}}}(A) = \text{R}\}$
$S :=$ rightSubtree.retrieveFrames (A)

Algorithm 2: MSER-LAF-TREE: Learning

Input: S_A: Set of LAFs of one type

Tree.build (S_A)
$S := \emptyset$
for all $A \in S_A$ **do**
$S := S \cup \{\{A, 1\}\}$ {assign unit probability}
root.build (S)

Node.build (S)
if $|S| \leq$ leaf capacity **or** indistinguishable (S) **then**
isLeaf := true, leafFrames := S
else
$d_{\overline{\mathbf{x}},\Theta_{\overline{\mathbf{x}}}} :=$ selectClassifier (S)
$S_L = \emptyset, S_R = \emptyset$
for all $\{A, p_A\} \in S$ **do**
$p_L := p_A \cdot p(d_{\overline{\mathbf{x}},\Theta_{\overline{\mathbf{x}}}}(r(A)) = \text{L})$
$p_R := p_A \cdot p(d_{\overline{\mathbf{x}},\Theta_{\overline{\mathbf{x}}}}(r(A)) = \text{R})$
if $p_L \geq \Theta_p$ **then**
$S_L := S_L \cup \{\{A, p_L\}\}$
if $p_R \geq \Theta_p$ **then**
$S_R := S_R \cup \{\{A, p_R\}\}$
if $S_L \neq \emptyset$ **then**
leftSubtree.build (S_L)
if $S_R \neq \emptyset$ **then**
rightSubtree.build (S_R)

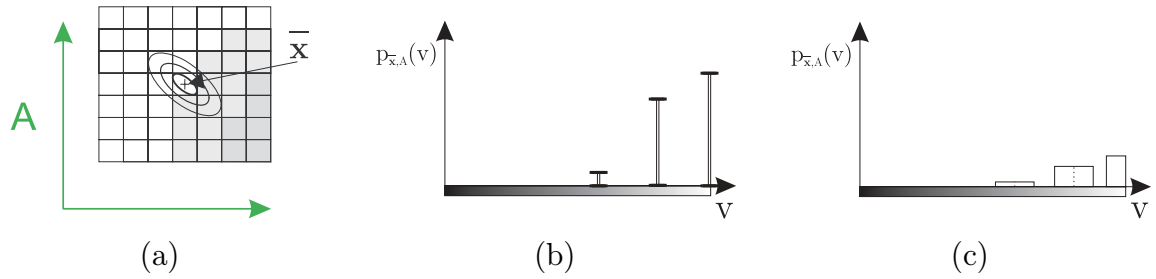


(a) (b) (c)

Figure 3: Geometric misalignment of detected frames, experimentally obtained for different types of frame constructions. The images show covariance matrices of distributions of displacements of pixels in rasterised patches, centred around detected LAFs. (a) LAF construction based on normalisation by region covariance matrix, (b) LAF construction based on a bi-tangent segment, (c) LAF construction based on normalisation by covariance matrix of a concavity [8]

2.3 Considering the estimated geometric uncertainty

. Having a database frame A of certain type, what is the probability of observing value v at measurement position $\overline{\mathbf{x}}$ in a corresponding query frame $r(A)$? The situation is depicted in Figure 4. Fig. 4(a) illustrates a part of the patch around measurement position $\overline{\mathbf{x}}$, and the



(a) (b) (c)

Figure 4: Probability of observing value v at position $\overline{\mathbf{x}}$ in a query realisation of frame A (a) localisation uncertainty $l_{\overline{\mathbf{x}}}$ for a pixel at position $\overline{\mathbf{x}}$, (b) probability $p_{\overline{\mathbf{x}},A}(v)$ of value v, (c) the probability after considering photometric noise

corresponding distribution of localisation uncertainty $l_{\overline{\mathbf{x}}}$ for that particular frame type. The probability $p(v)$ of observing a value v in a query frame at position $\overline{\mathbf{x}}$ is given as

$$p_{\overline{\mathbf{x}},A}(v) = \int_{\Omega_{v,A}} l_{\overline{\mathbf{x}}} \, \mathrm{d}\Omega, \tag{1}$$

where $\Omega_{v,A}$ is the area in A covered by pixels of value v. Fig. 4(b) shows the resulting distribution $p_{\overline{\mathbf{x}},A}(v)$ for the example from Fig. 4(a). Narrow distributions of $p_{\overline{\mathbf{x}},A}(v)$, which are benign for unambiguous decisions about query frames, are intuitively obtained either in areas of uniform intensity or where the localisation is precise.

The framework also consistently handles situations when some of the measurements are undefined, e.g. because not being on the object. Let us consider a case, where the outline of the model object is available (as in Figure 5(a)). Some of the frames will partially cover an area not on the object. In this area, the model cannot predict what value v will occur in a query frame. Since we do not build any background model (the probability distribution of intensities in the scene background) we consider all values v equiprobable. That is, if $\overline{\mathbf{x}}$ is outside of the object, $p_{\overline{\mathbf{x}},A}(v)$ has flat distribution over the whole domain of v ($p_{\overline{\mathbf{x}},A}(v) = 1/256$ for $v \in \{0, \ldots 255\}$).

2.4 Modelling photometric noise

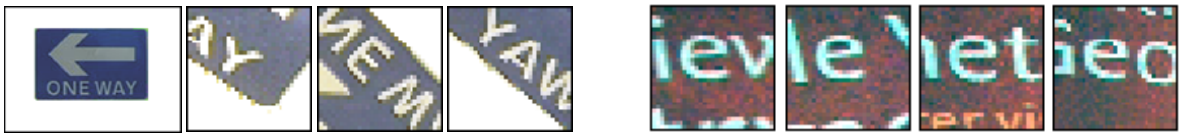
. A very simple model of photometric noise is employed – the noise distribution is assumed to be flat in a range of $(-\epsilon, \epsilon)$ intensity values. As illustrated in Figure 4(c), the probability of observing v becomes $p_{\overline{\mathbf{x}},A}(v)/\epsilon$. In the experiments, ϵ is set to 10, independently of v.

Going back to the Algorithm 2, the probabilities that a query realisation $r(A)$ will descend into the left and right subtree respectively are expressed as

$$p(d_{\overline{\mathbf{x}},\Theta_{\overline{\mathbf{x}}}}(r(A)) = \mathrm{L}) = \int_0^{\Theta_{\overline{\mathbf{x}}}} p_{\overline{\mathbf{x}},A}(v) \, \mathrm{d}v, \text{ resp. } p(d_{\overline{\mathbf{x}},\Theta_{\overline{\mathbf{x}}}}(r(A)) = \mathrm{R}) = \int_{\Theta_{\overline{\mathbf{x}}}}^{255} p_{\overline{\mathbf{x}},A}(v) \, \mathrm{d}v \tag{2}$$

for $v \in \{0 \ldots 255\}$.

The remaining issue in the tree construction algorithm is the choice of classifiers for non-terminal nodes. The objective is to minimise the expected recall time for query frames. Note that it does not necessarily mean to minimise the tree depth, since the distributions of database and query frames can significantly differ. if we would be able to construct a tree in which most background frames (which may easily account for 99% of all query frames) are discarded by first few decisions, the overall response time would be faster even if the foreground frames are deep in the tree. But the modelling of the background have



(a) (b)

Figure 5: The need for variable-sized measurement regions. (a) An example of a segmented model image and some of its frames. Using a common fixed measurement region where values are defined for all frames would lead to small nondescriminative descriptors. Large regions would include background in test images. (b) Frames detected on multiple instances of the 'e' letter on the 'Multiple view geometry' book title. The patches cannot be distinguished close to the detected frames and a distant measurement (e.g. on a neighbouring letter) is needed to separate them.

not been implemented yet, and in the following we assume that the distribution of query frames is close to that of the frames on the objects. The assumption holds for closed-world setups, e.g. for the COIL-100 database (see Section 3).

To select the classifier for a non-terminal node, let us have a set S of frames A, each with assigned probability p_A – the probability that $r(A)$ will descend from root to that node. The task is to select a measurement position $\overline{\mathbf{x}}$ and a threshold $\Theta_{\overline{\mathbf{x}}}$ so that, on average, the queries reach leaf nodes in minimal time, i.e. on minimal tree level. The requirements translate to (a) that the tree is balanced for query frames (thus, due to the closed-world assumption, for the stored frames), and (b) the number of ambiguous frames stored in *both* subtrees is minimised. It follows from (a) that for any particular $\overline{\mathbf{x}}$, the threshold $\Theta_{\overline{\mathbf{x}}}$ is set to median value, so that

$$\begin{aligned} \sum_{A\in S} p_A\, p(d_{\overline{\mathbf{x}},\Theta_{\overline{\mathbf{x}}}}(r(A)) = \mathrm{L}) &= \sum_{A\in S} p_A\, p(d_{\overline{\mathbf{x}},\Theta_{\overline{\mathbf{x}}}}(r(A)) = \mathrm{R}) \\ \sum_{A\in S} p_A \int_0^{\Theta_{\overline{\mathbf{x}}}} p_{\overline{\mathbf{x}},A}(v)\,\mathrm{d}v &= \sum_{A\in S} p_A \int_{\Theta_{\overline{\mathbf{x}}}}^{255} p_{\overline{\mathbf{x}},A}(v)\,\mathrm{d}v \end{aligned} \tag{3}$$

The measurement position $\overline{\mathbf{x}}$ that best separates (minimises overlap) of the frames in S is selected as

$$\overline{\mathbf{x}} = \operatorname*{argmin}_{\overline{\mathbf{x}}} \frac{1}{|S|} \sum_{A\in S} \min\Big(p_A\, p(d_{\overline{\mathbf{x}},\Theta_{\overline{\mathbf{x}}}}(r(A)) = \mathrm{L}), p_A\, p(d_{\overline{\mathbf{x}},\Theta_{\overline{\mathbf{x}}}}(r(A)) = \mathrm{R})\Big) \tag{4}$$

Ideally, when a position $\overline{\mathbf{x}}$ (and a corresponding threshold $\Theta_{\overline{\mathbf{x}}}$) is found which perfectly separates the set S, the minimised term evaluates to zero. In the worst case of identical distributions $p(d_{\overline{\mathbf{x}}}, \Theta_{\overline{\mathbf{x}}}(r(A)))$ for all $A \in S$, the term evaluates to 0.5. Let us consider the example shown in Figure 5 (b). In idealised case of noise-free images and perfect alignment of frames, no measurement positions $\overline{\mathbf{x}}$ on the letter 'e' nor the brown book background will allow for discrimination of the frames. The formula 4 ensures that rather a distant but discriminative measurement is selected.

3 Experiments

The performance of the proposed method, both in the recognition rate and execution speed, was evaluated on two datasets. The COIL-100 dataset has been widely used in

object recognition literature [12, 8, 4, 2, 13], and the experiment is included to compare the recognition rate with other state-of-the-art methods. ZuBuD, the second dataset, represents a larger, real-world problem, with images taken outdoor, with occluded objects, varying background, and illumination changes[3].

(a) (b)

Figure 6: COIL-100 [1]: (a) Several objects from the database, (b) Examples of query images for the occlusion experiment

(a) (b)

Figure 7: ZuBuD dataset [10]: Examples of (a) query and (b) the corresponding database images.

3.1 COIL-100.

The Columbia Object Image Library (COIL-100) [1] is a database of colour images of 100 different objects; 72 images of each object placed on a turntable were acquired at pose intervals of 5°. Neither occlusion, background clutter, nor illumination changes are present. Several images from the database are shown in Figure 6(a). First two experiments were performed that differ in the number of images used for training. The achieved recognition rate was 98.2% for 4 training views per object (90° apart, 68 test views per object) and 99.7% for 8 training views (45° apart, 64 test views). Table 1 summarises the results and provides comparison to other published results.

In another experiment, occlusion of the objects was simulated by blanking one half of the test images (see Figure 6 (b)). Four full (unoccluded) training views per object were used in training. The recognition rate was 87%, which is comparable to the results of other methods on unoccluded images.

Table 2 provides qualitative information about the experiments. Two variants of the recognition system were evaluated, one which recalls the stored frames via the proposed decision tree (with sublinear recall time), and a second one which sequentially scans through all stored frames (linear recall time). Note that doubling the number of training images (columns 2 and 3) did not double the recall time for the tree approach. This confirms the claim that the recall time is sub-linear in the number of stored frames. The recall times in the Table 2 show that using the decision tree matching of approximately 500

[3]Many ad-hoc experiments have been performed using a live recognition system. The system is capable of recognition of about 100 of objects with response time under 1 second, is viewpoint insensitive, and handles partially occluded objects and cluttered background. Demonstration of the system will be shown during the conference.

query frames against hundreds of thousands of stored frames takes about 2 milliseconds. The total response time of the recognition system is the sum of the time needed to build the query image representation (independent of the number of database objects – 7th row of Table 2) and the recall time (8th or 9th row).

Method	8 views	4 views	Method	8 views	4 views
MSER+LAF+tree (proposed)	99.8%	98.2%			
MSER+LAF 2002 [8]	99.4%	94.7%	Spectral representation [4]	96.3%	–
Kullback-Leibler SVM [12]	95.2%	84.3%	SNoW / edges [13]	89.2%	88.3%
Spin-Glass MRF [2]	88.2%	69.4%	SNoW / intensity [13]	85.1%	81.5%
Linear SVM [13]	84.8%	78.5%	Nearest Neighbour [13]	79.5%	74.6%

Table 1: COIL-100 experiment: Comparison with published results

	COIL-100			ZuBuD
Occluded queries	no	no	yes	n/a
Training view dist	90°	45°	90°	n/a
Number of DB images	400	800	400	1005
Number of DB frames	186346	385197	186346	251633
Number of query images	6800	6400	6800	115
Avg number of query frames	494	494	269	1594
avg time to build representation	520 ms	522 ms	251 ms	1255 ms
avg recall time without the tree	493 ms	3471 ms	277 ms	27234 ms
avg recall time with the tree	1.99 ms	2.17 ms	1.07 ms	14.3 ms
recognition rate	98.24%	99.77%	87.01%	93 %

Table 2: Experimental results on COIL-100 and ZuBuD datasets

3.2 ZuBuD dataset

. The experiment was conducted on a set of images of 201 buildings in Zurich, Switzerland, which is publicly available [10]. The database consists of five photographs of every of the 201 buildings. A separate set of 115 query images is provided. For every query image, there are exactly five matching images of the same building in the database. Query and database images differ in viewpoint, variations in the illumination are present, but rare. Examples of corresponding query and database images are shown in Figure 7. Experimental results are summarised in the last column of Table 2. The slower recall times, compared with the COIL-100 dataset, are caused by a higher number of query frames and by the increase of the leaf capacity from 4 to 10 – up to 10 frames were searched exhaustively in the leaf nodes. The leaf capacity represents a trade-off between recall speed and recognition rate. Setting the capacity to 1000, a recognition rate of 98.2% was achieved, but the average recall time dropped to 510 ms. Linear exhaustive scan through all the stored frames (avoiding the tree) achieved recognition rate of 100% [9], but with recall times over 27 seconds per image.

4 Conclusions

An object recognition method capable of sub-linear recall has been proposed. Objects are represented by local affine frames, i.e. as a set of local photometric measurements expressed

in object-centred coordinates. The local affine frames are stored in a binary decision-measurement tree organised to minimise average time to decision. A frame is recognised by descending the tree where each decision not only reduces the number of potential corresponding frames represented in the tree but also defines which measurements are taken next.

We show experimentally that the method supports near real-time recognition of hundreds of real-world objects with state-of-the-art recognition rates. Establishing correspondences between hundreds of query local frames and hundreds of thousands of stored frames takes only a few milliseconds. The proposed LAF-TREE method possesses all the strengths of local region methods – robustness to background clutter, occlusion, and large changes of viewpoints.

References

[1] Columbia object image library. http://www.cs.columbia.edu/CAVE.

[2] B. Caputo, J. Hornegger, D. Paulus, and H. Niemann. A spin-glass markov random field for 3-D object recognition. Technical Report LME-TR-2002-01, Lehrstuhl für Mustererkennung, Institut für Informatik, Universität Erlangen-Nürnberg, 2002.

[3] V. Lepetit, P. Lagger, and P. Fua. Randomized trees for real-time keypoint recognition. In *Computer Vision and Pattern Recognition*, June 2005.

[4] X. Liu and A. Srivastava. A spectral representation for appearance-based classification and recognition. In *ICPR (1)*, pages 37–40, 2002.

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

[6] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In P. L. Rosin and D. Marshall, editors, *Proceedings of the British Machine Vision Conference*, volume 1, pages 384–393, London, UK, September 2002. BMVA.

[7] S. A. Nene and S. K. Nayar. A simple algorithm for nearest neighbor search in high dimensions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(9):989–1003, 1997.

[8] Š. Obdržálek and J. Matas. Object recognition using local affine frames on distinguished regions. In *The British Machine Vision Conference (BMVC02)*, September 2002.

[9] Š. Obdržálek and J. Matas. Image retrieval using local compact dct-based representation. In *DAGM 2003: Proceedings of the 25th DAGM Symposium*, pages 490–497, 9 2003.

[10] H. Shao, T. Svoboda, and L. Van Gool. ZuBuD — Zurich Buildings Database for Image Based Recognition. Technical Report 260, Computer Vision Laboratory, Swiss Federal Institute of Technology, March 2003. http://www.vision.ee.ethz.ch/showroom/zubud.

[11] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV03*, pages 1470–1477, 2003.

[12] N. Vasconcelos, P. Ho, and P. J. Moreno. The Kullback-Leibler kernel as a framework for discriminant and localized representations for visual recognition. In *ECCV (3)*, pages 430–441, 2004.

[13] M. H. Yang, D. Roth, and N. Ahuja. Learning to Recognize 3D Objects with SNoW. In *ECCV 2000*, pages 439–454, 2000.

5 Dr. Ing. Jiří Matas

Vzdělání.

1995 PhD, University of Surrey, Velká Británie.
1987 Ing. (s vyznamenáním) v oboru technická kybernetika.

Zaměstnání.

2005-2006 Výzkumný pracovník, Centre for Vision, Speech and Signal Processing, University of Surrey, Velká Británie.
1997-2005 Výzkumný pracovník, Centrum strojového vnímání, Katedra kybernetiky, FEL ČVUT Praha.
1991-1997 Výzkumný pracovník, Centre for Vision, Speech and Signal Processing, University of Surrey, Velká Británie.
1987-1991 Aspirant, Katedra řídící techniky, FEL ČVUT Praha.

Ocenění.

2005 Vedl tým, který obsadil 2. místo v soutěži v automatické lokalizaci ICCV 2005 Contest. ICCV, International Conference on Computer Vision, je nejprestižnější konferencí v oboru.
2005 Cena za nejlepší článek na British Machine Vision Conference v Oxfordu, Velká Británie.
2004 Cena rektora ČVUT 1. stupně za vynikající vědecky výsledek.
2002 Cena za nejlepší článek na British Machine Vision Conference v Bristolu, Velká Británie.

Projekty.

V letech 1992-2006 pracoval na projektech EU, a to projektu základního výzkumu EU BRA 3038 VAP "Vision as Process" (1991-1995), EU ACTS projektu M2VTS "Multimodal Verification for Teleservices and Security Applications" (1995-1997), na projektech 5. rámcového programu EU BANCA "Biometric access control for networked and e-commerce applications" (1997-1999) a na projektu ACTIPRET "Activity Interpretation" (2001-2004). Od roku 2004 pracuje na projektu EU COSPAL "COgnitive Systems using Perception-Action Learning" Byl řešitelem grantu GAČR "Rozpoznávání tváří (za libovolných podmínek) z jediného snímku" (2001-2004).

Výuka, vedení doktorandů.

Přednáší předměty magisterského studia "Digitální zpracování obrazu" a "Rozpoznávání", a doktorandský předmět "Vybrané partie z rozpoznávání". V roce 2005 přednášel v kurzu "Image processing and vision" na University of Surrey. V letech 1996-2003 se podílel na vedení 4 PhD studentů na University of Surrey ve Velké Británii. Všichni úspešně studium dokončili. Na ČVUT úspěšně vedl 1 doktoranda (PhD práce obhájena v roce 2005). V současné době vede 4 doktorandy, z nichž 2 jsou před odevzdáním disertační práce.

Ostatní profesionální aktivity.

Je členem programového výboru řady prestižních konferencí: International Conference on Computer Vision, International Conference on Pattern Recognition, Computer Vison and Pattern Recognition, Neural Information Processing Systems, International Conference on Face and Gesture Recognition, Audio- and Video-based Biometric Person Authentication, International Conference on Image and Video Retrieval, British Machine Vision Conference.

J. Matas byl spolupředsedou programového výboru nejprestižnější konference v oboru počítačového vidění pořádaného v Evropě - European Conference on Computer Vision 2004.

J. Matas se stal v roce 2003 zástupcem ČR v ISOIEC JTC 1SC 29WG 11(MPEG).

Návrh standardu reprezentace obličeje, na kterém se J. Matas podílel, se stal součásti normy MPEG-ISO.

J. Matas byl v letech 2002-2004 předsedou technické komise TC 14 "Signal Processing for Machine Intellingence" mezinárodní organizace International Association for Pattern Recognition (od roku 2004 je místopředsedou této komise).

Patenty.

J. Matas je spoluautorem ("inventor") dvou patentových přihlášek.

Průmyslové aplikace.

J. Matas vedl řadu projektů aplikující výsledky v oblasti rozpoznávání do praxe. Vybrané projekty (zadavatel, trvání, náplň):

Hitachi, Japonsko	2004-2006	Vedoucí projektu. Detekce obličejů.
Toyota, Japonsko	2003-2007	Vedoucí projektu. Aplikace rozpoznávání objektů v dopravě.
Samsung, Jižní Korea	2001-2004	Vedoucí projektu. Detekce, verifikace identity a rozpoznávání obličejů.
Výzkukmný ústav letectví Praha	2001-2002	Vedoucí projektu. Rozpoznávání objektů ve videu z bezpilotního letounu.
Boeing , USA	1999-2000	Vývoj algoritmů pro rozpoznávání letadel.
Racal, Velká Británie	1996	Konzultant v projektu zabývajícím se rozpoznávaním poznávacích značek v USA.

Publikace, Citace v Science Citation Index.

Publikoval více než 130 článků v recenzovaných časopisech a konferenčních sbornících. Tyto články byly dohromady citovány více než 500 krát (bez přímých a nepřímých autocitací).