# Learning visual distance function for object identification from one example

Eric Nowak, Frédéric Jurie

# Learning Visual Distance Function for Object Identification from one Example

**Eric Nowak**
INPG - INRIA - Bertin Technologies
155, r Louis Armand
13290 Aix en Provence
eric.nowak@inrialpes.fr

**Frédéric Jurie**
CNRS - INRIA
655 avenue de l'Europe
38334 Saint Ismier Cedex France
frederic.jurie@inrialpes.fr

## 1 Introduction

Comparing images is essential to several computer vision problems, like image retrieval or object identification. The comparison of two images heavily relies on the definition of a good distance function. Standard functions (e.g. the euclidean distance in the original feature space) are too generic and fail to encode the domain specific information.

In this paper, we propose to learn a similarity measure specific to a given category (e.g. cars). This distance is learned from a training set of pairs of images labeled "same" or "different", indicating if the two images represent the same object (e.g. same car model) or not. After learning, this measure is used to predict how similar two images of *never seen* objects are (see figure 1).

Most of the contributions to solve this problem are inspired by the Mahanalobis distance [9, 5, 1]. One can also model expected deformations of object appearances[7, 3, 6]. Some authors explicitly deal with robustness to change in pose, illumination condition, clutter [2]. [2] encourages us to adapt our previous work [8] based on a bag-of-words model and an ensemble of extremely-randomized trees.

In section 2 we give an overview of the approach and in section 3 we show experimental results.

## 2 Building a similarity measure from patch correspondences.

Our objective is to build a similarity measure for deciding whether two images represent the same object instance or not, and it is trained from pairs of "same" and "different" objects of the same generic category (e.g. cars), without knowing the object precise categories (e.g. the car model in each image).

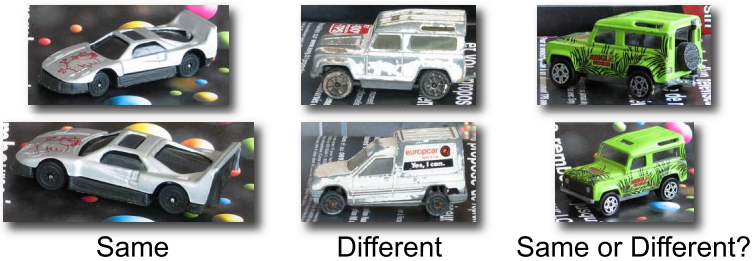


Figure 1: Is it possible to learn the notion of same/different object on a training set of pairs of images, and to design a similarity measure that predicts if two images represent the same *never seen* object?

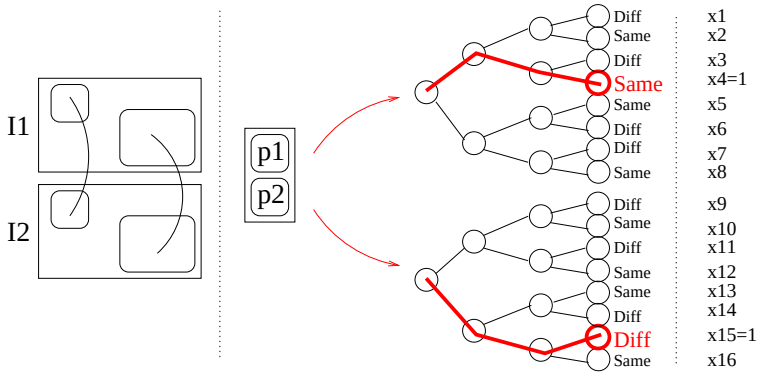


Figure 2: Similarity computation. Left: Corresponding patch pairs from both images. Middle: Assign each patch pair to a cluster (a leaf) with the set of trees. Right: Produce the list of reached pairs $f(p_1, p_2) = (0, 0, 0, 1, \ldots, 1, 0)$.

The computation of the measure is a three step process. (a) Several pairs of corresponding local regions (patches) are sampled from a pair of images. (b) Each pair is assigned to a cluster by an ensemble of extremely randomized decision trees. Although the trees are trained to separate "same" and "different" patch pairs, we discard their decision ("same", "different") and instead we consider which leaves are reached by the patch pairs. (c) The clusters are combined to make a global decision about the pair of images. Those steps are detailed below.

**Sampling patch pairs.** Each patch pairs is produced as follow. A patch $p_1$ of a random size is cropped at a random position $(x, y)$ in the first image $I_1$. The best normalized cross correlation match $p_2$ is looked for in the second image $I_2$, in the neighborhood of $(x, y)$ (for example a region twice bigger than the patch). $p_1$ and $p_2$ are then resized to a standard size $w_{std}$ so that all patches are comparable and scale independent.

**Learning an ensemble of extremely randomized trees.** All trees are learned independently, according to the following procedure. We sample a large number of patch pairs from positive image pairs (same object) and negative image pairs (different objects). We then create a tree with a unique node, the root node, that contains all these pairs. We recursively affect a split condition to a node and then split it into sub-nodes until the nodes contain only positive or negative patch pairs. The split conditions are created randomly, and they consist in very simple tests on pixel intensity such as: "Are the gray level values at position (8,6) in the two patches larger than 0.8?".

**Assigning patch pairs to clusters.** Patch pairs are assigned to a cluster via the learned ensemble of trees, the leaves being the clusters. Each patch pair is input in the root node of all trees. For each tree, the patch pair goes from the root node to a leaf (see figure 2, middle), at each node the left or right child node is selected according to the evaluation of the split condition of that node. When a patch pair reaches a leaf, the corresponding leaf label (i.e. the id of the leaf in the forest) is set to 1. If a leaf is never reached, it is set to 0. Thus, an image pair is transformed into a binary feature vector (of size the total number of leaves), each dimension indicating if a patch pair sampled from the image pair has reached the corresponding leaf (see figure 2, right).

**Computation of the similarity measure.** The similarity measure is a simple linear combination of the elements of the binary feature vector. The weights are optimized such as the higher the similarity measure, the more confidence in the similarity of the two images. In practice, the weight vector is the hyperplane normal of a binary linear SVM trained on the binary feature vectors of positive and negative image pairs.

## 3 Experimental results.

We evaluate our similarity measure on 3 different datasets: a small dataset of toy cars[1] and two other publicly available datasets[2] making comparison with competitive approaches possible. For each dataset, images are aligned and we have pairs marked as positive (same object) or negative (different objects). Those sets are split into a training set and a test set. All the objects of the test set are new, they were never seen during training. Thus, the similarity measure is evaluated on *never seen* objects. We compute a Precision-Recall Equal Error Rate score on the similarity measure of the test set image pairs to evaluate our performance.

**Comparison with state-of-the-art competitive approaches.** On the toy car dataset, we get an EER-PR of 79.7% $\pm$ 0.0 on 5 runs. On the Ferencz & Malik dataset, we get an EER-PR of 87.7%$\pm$0.7 on 5 runs, where Ferencz & Malik get 84.9%. On the Coil-100 dataset, we have a misclassification rate of 11.7$\pm$3 where Fleuret & Blanchard have 11.4%$\pm$4 (comparable given the high variance). However, the method of Fleuret and Blanchard uses the information of the real object categories during training, whereas we only know if two images belong to the same category or not.

## 4 Conclusions and future works

We are dealing with the problem of predicting how similar two images of never seen objects are given a training set of similar and different object pairs. We proposed an original method consisting in (a) clustering a set of corresponding local regions sampled in the two images by an ensemble of randomized trees and (b) combining the cluster membership of the pair of local regions to take a global decision about the two images. Our experiments show that our approach gives good results on the three publicly available datasets we have used for evaluation.

We are currently extending our approach to non aligned objects, that is the object of interest may be anywhere in the input image. This problem is much more challenging because it can hardly rely on the search for the corresponding local regions.

## References


[1] A. Bar Hillel, T. Hertz, N. Shental, and D. Weinshall. Learning a mahalanobis metric from equivalence constraints. *Journal of Machine Learning Research*, (6):937–965, 2005.

[2] A. Ferencz, E.G. Learned Miller, and J. Malik. Building a classification cascade for visual identification from one example. In *ICCV'05*, pages I: 286–293, 2005.

[3] A.W. Fitzgibbon and A. Zisserman. Joint manifold distance: a new approach to appearance based clustering. In *CVPR'03*, pages I: 26–33, 2003.

[4] F. Fleuret and G. Blanchard. Pattern recognition from one example by chopping. In *NIPS'05*, pages 371–378. MIT Press, 2005.

[5] A. Globerson and S. Roweis. Metric learning by collapsing classes. In *NIPS'05*, 2005.

[6] F.F. li, R. Fergus, and P. Perona. One-shot learning of object categories. *PAMI*, 28(4):594–611, April 2006.

[7] E.G. Miller, N.E. Matsakis, and P.A. Viola. Learning from one example through shared densities on transforms. In *CVPR'00*, pages I: 464–471, 2000.

[8] F. Moosmann, B. Triggs, and F. Jurie. Randomized clustering forests for building fast and discriminative visual vocabularies. In *NIPS'06*. 2006.

[9] K. Weinberger, J. Blitzer, and L. Saul. Distance metric learning for large margin nearest neighbor classification. In *NIPS'05*. 2006.


[1] publicly available at http://lear.inrialpes.fr/people/nowak/dwl/toycarlear.tar.gz

[2] the cars of Ferencz & Malik [2] and the Coil-100 dataset of Fleuret & Blanchard [4]