Multimodal Object Recognition using Random Clustering Trees

M. Villamizar, A. Garrell, A. Sanfeliu and F. Moreno-Noguer

{mvillami,agarrell,sanfeliu,fmoreno}@iri.upc.edu
Institut de Robotica i informatica industrial CSIC-UPC
Barcelona - Spain

Abstract. In this paper, we present an object recognition approach that in addition allows to discover intra-class modalities exhibiting high-correlated visual information. Unlike to more conventional approaches based on computing multiple specialized classifiers, the proposed approach combines a single classifier, Boosted Random Ferns (BRFs), with probabilistic Latent Semantic Analysis (pLSA) in order to recognize an object class and to find automatically the most prominent intra-class appearance modalities (clusters) through tree-structured visual words. The proposed approach has been validated in synthetic and real experiments where we show that the method is able to recognize objects with multiple appearances.

Keywords: object recognition, random trees, clustering, boosting

1 Introduction

Computer vision is nowadays a very active research field where it has made great strides in recent years, especially in the recognition of objects in images and videos. Currently, there exist methods that can detect and identify objects with outstanding results despite the large difficulties present in this problem such as intra-class variations, 3D rotations, scaling, illumination changes [2,4,8,14].

However, most of these methods are based on complex algorithms that depend of a rigorous training and large object databases. Usually, these methods compute object detectors using a supervised and offline learning where time constraints and computational cost are not a big issue.

In order to compute efficient and robust object detectors, approaches based on randomized trees have been proposed in the past with outstanding results, especially in terms of efficiency and reliability [9,7]. Particularly, these methods have been focused mainly on the fast matching of binary descriptors. Subsequently, a robust and efficient classifier for the detection of object classes was proposed in [16]. This method, called Boosted Random Ferns (BRFs), combines multiple extremely randomized trees (e.g Random Ferns [9]) using AdaBoost so as to select automatically the most relevant trees in one single classifier.

Although this classifier has shown remarkable results to detect objects with multiple intra-class modalities (e.g multiple object's views), this method is unable to distinguish these modes automatically. For this purpose, methods based

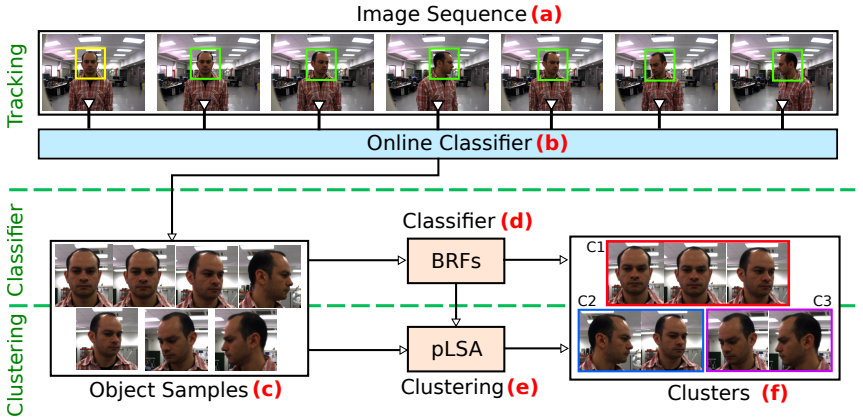

Fig. 1. Overall scheme of the proposed approach to compute object classifiers using weakly supervised learning. In this approach only the first frame is annotated manually. For clarity, this figure does not include background (negative) samples.

on the computation of multiple specialized classifiers have been proposed, where each one is devoted to a particular appearance cluster [6,13]. However, these methods increase the complexity and computational cost of the detector since various classifiers are considered during run time. Additionally, computing these classifiers in a supervised learning require annotating all training samples with their corresponding appearance cluster. This task is cumbersome and tedious since it is usually carried out manually.

In this work, we present a more straightforward approach to recognize object appearance clusters (i.e, intra-class modalities) using weakly human supervision during the training phase. More precisely, the proposed method consists of three main stages, observe Fig. 1. In the initial stage (*object tracking*), an online classifier is computed in order to detect and track the object through a sequence of images (Fig. 1-a,b). This process is automatic and requires only the assignment of the object in the first frame using a bounding box (yellow box). The result of this step is a set of training samples (images) of the object with different appearance (Fig. 1-c). In the second stage (*classifier*), a more robust classifier (BRFs) is computed using the training samples (Fig. 1-d). Finally, in the third step (*clustering*), the training samples are clustered using probabilistic Latent Semantic Analysis (pLSA) [1,5,11] and the responses of the BRFs classifier on the samples (Fig. 1-e). Fig. 1-f shows as example three clusters of training samples grouped according their visual similarity.

The method we present is a further step of the approach proposed in [3,12] for learning and detecting objects using human-robot interactions. Actually, this method corresponds to the *tracking* stage in Fig. 1. In this work, we combine this method with BRFs and pLSA in order to detect and distinguish multiple object appearances. This is particular useful for robotics applications where knowing a specific object view allows to take actions. For example, for human-robot interaction is important to determine whether people look at the robot (see Fig. 1).

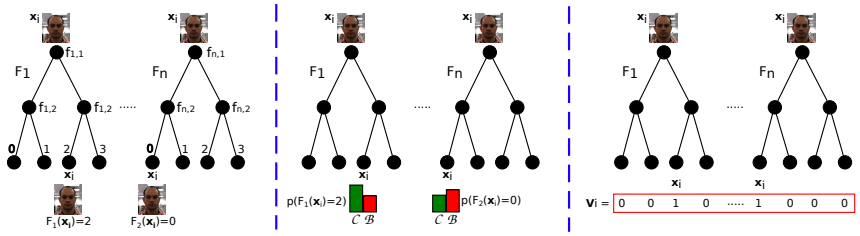

Fig. 2. Online classifier used to detect and track the object trough an image sequence.

2 Proposed Approach

2.1 Object Tracking

The first stage of the proposed method corresponds to perform object detection and tracking over an input image sequence, observe Fig. 1. The goal of this stage is to extract automatically a set of training samples which are used later to compute the object classifier.

To track the object in every frame, we compute an online classifier based on extremely randomized trees [9,12]. This classifier is initialized using an object annotation provided by the user in the first frame (Fig. 1-a). Subsequently, the classifier is computed and updated incrementally using their own detection hypotheses on new input images. This self-learning approach allows computing and adapting an object detector while discovers object instances in images.

More formally, the classifier is comprised of a series of N random ferns where each fern F_k computes a set of M signed comparisons between pairs of intensity pixel values $\{f_{k,m}\}_{m=1}^{M}$, known commonly as binary features. Fig. 2-left shows for example the output of two fern instances over an input sample $\mathbf{x}_i$. We observe that the fern output $F_k(\mathbf{x}_i)$ depends of the responses of the binary features. The co-occurrence of these features determines the tree leaf where the sample falls.

Once the response of each fern k is computed $F_k(\mathbf{x}_i)$, the classifier updates its class-conditional probabilities in each tree, $p(F_k(\mathbf{x}_i)|\mathcal{C})$ and $p(F_k(\mathbf{x}_i)|\mathcal{B})$, according whether the sample $\mathbf{x}_i$ belongs to either the object $\mathcal{C}$ or background $\mathcal{B}$ class. This is illustrated in Fig. 2-middle, where the input sample is used to update the fern distributions. For further information about this online classifier and its computation see [12,15].

2.2 The Object Classifier

The object classifier is computed using Boosted Random Ferns (BRFs) since they have demonstrated to be an efficient and robust classifier for object recognition [16]. Further in detail, the object classifier $H(\mathbf{x})$ is built using a boosting combination of weak classifiers h_t where each is a random fern F_t computed to particular object location (u_t, v_t). The classifier is computed in order to find the ferns and locations that most discriminate the object (positive) class from the background (negative) one. The computation of the classifier is done using real AdaBoost, that iteratively assembles weak classifiers and adapts their weighting values to focus all its effort on the misclassified samples from previous weak classifiers [10].

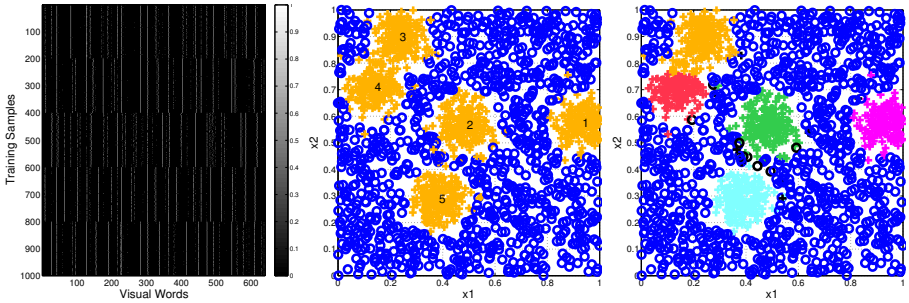

Fig. 3. Computation of object class clusters using pLSA. Left: table including the co-occurrence between training samples and visual words. Middle: two-class classification problem in 2D feature space. The positive class (crosses) has five intra-class modalities. Right: output of the clustering stage to find latent topics (sample clusters).

The object classifier with T weak classifiers is then defined as:

$$H(\mathbf{x}) = \sum_{t=1}^{T} h_t(\mathbf{x}) > \beta \ , \tag{1}$$

where $\mathbf{x}$ is a test sample, β is the classifier threshold and h_t is a weak classifier computed by

$$h_t(\mathbf{x}) = \frac{1}{2} \log \frac{p(F_t(\mathbf{x}) = r|\mathcal{C}) + \epsilon}{p(F_t(\mathbf{x}) = r|\mathcal{B}) + \epsilon} \ , \tag{2}$$

where r is the output of the fern F_t on the sample $\mathbf{x}$, ϵ is a smoothing parameter, and $\mathcal{B}$ and $\mathcal{C}$ are the background and object class labels, respectively. In order to extract the most discriminative weak classifier at each iteration, the AdaBoost algorithm seeks for the fern that minimizes the distance between class-conditional probabilities, $p(F_t(\mathbf{x})|\mathcal{C})$ and $p(F_t(\mathbf{x})|\mathcal{B})$. For more information refer to [16].

2.3 Clustering

With the aim of finding important internal structures of the object class without human supervision, we propose to use pLSA to discretize the overall appearance of the object in multiple clusters of samples with a strong feature similarity. More specifically, pLSA is a generative model from the statistical text literature that allows discovering latent variables (topics) from a corpus containing co-occurrences between documents and words [1,5,11].

In this work, we use image samples and tree-structured visual words instead of text documents and words in order to find the most relevant clusters of the object appearance (topics). The pLSA algorithm is suitable for this problem because it provides a statistical model that allows represent an object sample $\mathbf{x}_i$ as a mixture of K topics,

$$p(w_j|\mathbf{x}_i) = \sum_{k=1}^{K} p(z_k|\mathbf{x}_i) p(w_j|z_k), \tag{3}$$

where $p(z_k|\mathbf{x}_i)$ is the probability of topic z_k occurring in the sample $\mathbf{x}_i$ whereas $p(w_j|z_k)$ is the probability of the visual word w_j occurring in the topic z_k [11].

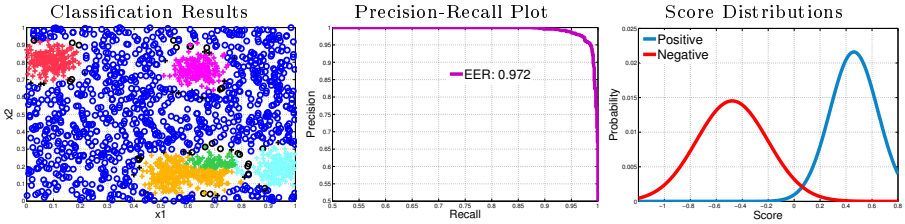

Fig. 4. Left: 2D classification results provided by the BRFs classifier. Crosses are positive samples while circles indicate negative ones. Black samples correspond to misclassified samples. Middle: classification plots using recall-precision curves and equal error rate (EER). Right: class score distributions for the positive and negative classes.

The pLSA computation is done using the EM algorithm and an input table containing the co-occurrence of training samples and the bag of visual words. Fig. 3-left shows this table where the object samples have been ordered by cluster in order to distinguish visually the strong patterns in the corpus. For our case, we use the object samples extracted by the online classifier, and define that each fern leaf j corresponds to a particular visual word w_j since it represents a specific configuration of binary features. This is shown in Fig. 2-right. For this simple example, the activated visual words (i.e, ones) co-occur with the input sample $\mathbf{x}_i$ since this sample falls in the corresponding fern leaves.

Finally, in Fig. 3-right we show the output of the clustering stage over a set of training samples using a 2D feature space with complex and multimodal class distributions (see Fig. 3-middle). In this figure, crosses are object or positive samples whereas circles make reference to background or negative samples. As a result, we can see that the positive samples are clustered in K=5 different clusters, indicated through different colors, and that each one keeps strong feature correlation in the 2D space.

3 Experiments

3.1 Synthetic Experiments - 2D Classification Problem

The proposed approach has been evaluated in synthetic experiments in order to observe more clearly the performance of the method. Fig. 4 shows, for example, the output of the proposed method on a scenario generated at random in which two class distributions (positive and negative) with high complexity are considered. For this experiment, the method was computed using $K = 5$ clusters.

We see in Fig. 4-left that the proposed approach achieves correctly classify most samples while discovers multiple intra-class modalities (indicated through clusters with different colors). The method only produces a small number of misclassified samples (black ones). This result is also shown in the precision-recall curve (Fig. 4-middle) where the method obtains a high equal error rate (EER). Moreover, the approach also increases the separability between classes and reduces the risk of misclassification.This is observed in Fig. 4-right where the class score distributions are shown.

The table 1 shows the average classification results of the BRFs classifier over 10 runs in order to consider the randomness of the classifier and the 2D scenario.

Classification Performance														
	RFs			BRFs										
	# Clusters (K)			# Clusters (K)			# Ferns (R)				# Features (S)			
	3	5	10	3	5	10	5	10	20	50	1	3	5	7
EER(%)	90.6	92.4	84.4	96.9	97.4	96.0	95.4	96.6	96.9	97.2	83.1	96.1	96.9	97.2
Distance(%)	59.0	68.8	46.5	83.1	90.6	73.2	77.2	80.4	83.1	88.8	41.2	73.0	83.1	88.4

Table 1. 2D classification results of the BRFs and RFs classifiers.

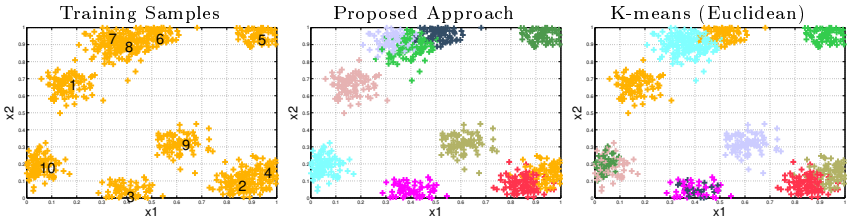

Fig. 5. 2D clustering results.

Each scenario is generated at random with multiple sample clusters (K). We see that BRFs obtain high classification rates (EER) and large distances between classes when the amount of features (F) and ferns (R) gets larger. Here, we use the Hellinger distance to measure the separability between Gaussian distributions. The default parameters for this experiment are $K = 5$, $R = 20$ and $S = 5$. The table also shows a comparison, in terms of the number of clusters, of BRFs against its counterpart without using boosting (RFs). Observe that the BRFs classifier attains the best performance rates.

Fig. 5 shows the clustering results of the presented approach in comparison with the K-means algorithm. The left figure corresponds to the positive training samples belonging to 20-dimensional feature space. In this figure, we only plot the first two feature dimensions ($x1$,$x2$). Fig. 5-middle plots the clustering output of the proposed approach (BRFs+pLSA), whereas the right figure shows the results of K-means using Euclidean distance in the sample feature space. We can see that our approach yields good clustering results, in contrast to the K-means algorithm which produces some incorrect clustering labels (observed through the confusion of colors in clusters). Finally, table 2 shows the average confusion values in the clustering labels for varying numbers of clusters and feature space dimensions (D). Here, we use as measure of confusion the entropy function over the confusion matrix (using ground truth labels). The table also includes a BRFs+K-means approach using the Hamming distance. As a result, we observe that the proposed approach (BRFs+pLSA) produces low confusion values, especially for large feature spaces.

3.2 Real Experiments - Multi-view Face and Object Detection

The proposed approach has also been tested to detect faces under multiple views, see Fig. 7. This corresponds to a classification problem involving multiple intra-classes where each one is associated to a particular view. For this experiment, we have used two face sequences of the dataset proposed in [12], where each sequence contains more than 200 images. For training, we have used the first sequence, whereas the second one is used for validation.

Clustering Results												
	BRFs+pLSA				K-means (Euclidean)				BRFs+K-means			
D	2	5	10	20	2	5	10	20	2	5	10	20
K=3	0.097	0.001	0.033	0.000	0.147	0.100	0.133	0.067	0.177	0.036	0.086	0.113
K=5	0.240	0.022	0.019	0.020	0.180	0.139	0.163	0.201	0.304	0.127	0.114	0.173
K=10	0.514	0.144	0.092	0.096	0.367	0.143	0.159	0.116	0.548	0.251	0.207	0.102

Table 2. Clustering results.

Fig. 6. Sample images showing three different face appearance clusters. Observe that samples belonging to the same cluster share visual similarities.

Fig. 6 shows some samples images corresponding to $K = 3$ different intra-class appearances modalities found by the proposed method during the training phase. We see that these samples share similar visual features and that the proposed method is able to cluster these samples using the output of a tree-structured classifier. In Fig. 7 are shown some detection results on the test images. Note that the method is capable of detecting most faces at the same time that it can estimate the face pose. This is indicated in the images through colored boxes. This experiment reveals that the proposed method using a single classifier can be used for pose estimation using the co-occurrence of visual words.

Similar to the previous experiment, our method has been tested for object recognition. In this case, for detecting a toy car from multiple viewpoints using $K = 5$ appearance clusters. Fig. 8 shows some example images where the response of the classifier is indicated by the bounding boxes and the color represents the object cluster. We can see that the proposed method is able to discretize automatically the overall object appearance in diverse modalities, each one corresponding to a particular object view.

4 Conclusions

In this paper, a weakly supervised learning approach has been proposed in order to compute an object classifier that is able to identify multiple intra-class modalities. The proposed approach combines a tree-structured classifier with a text document analysis algorithm so as to cluster the output of the classifier. The approach has been validated in synthetic and real experiments.

Acknowledgments Work partially supported by the Spanish Ministry of Science and Innovation under project DPI2013-42458-P, ERA-Net Chistera project ViSen PCIN-2013-047, and by the EU project ARCAS FP7-ICT-2011-28761.

References

1. A. Bosch, A. Zisserman, and X. Muñoz. Scene classification via plsa. In *ECCV*. 2006.
2. P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 32(9):1627–1645, 2010.

Fig. 7. Face detection results provided by the proposed approach. The output of the classifier is indicated by boxes whereas the face pose is shown through different colors.

Fig. 8. Object recognition results.

3. A. Garrell, M. Villamizar, F. Moreno-Noguer, and A. Sanfeliu. Proactive behavior of an autonomous mobile robot for human-assisted learning. In *RO-MAN*, 2013.
4. S. Hinterstoisser, V. Lepetit, P. Fua, and N. Navab. Dominant orientation templates for real-time detection of texture-less objects. In *CVPR*, 2010.
5. T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine learning*, 42(1-2):177–196, 2001.
6. T. K. Kim and R. Cipolla. Mcboost: Multiple classifier boosting for perceptual co-clustering of images and visual features. In *NIPS*, pages 841–848, 2009.
7. V. Lepetit and P. Fua. Keypoint recognition using randomized trees. *PAMI*, 28(9):1465–1479, 2006.
8. D.G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.
9. M. Ozuysal, M. Calonder, V. Lepetit, and P. Fua. Fast keypoint recognition using random ferns. *PAMI*, 32(3):448–461, 2010.
10. Robert E Schapire and Yoram Singer. Improved boosting algorithms using confidence-rated predictions. *Machine learning*, 37(3):297–336, 1999.
11. J. Sivic, B. Russell, A. Efros, A. Zisserman, and W.T. Freeman. Discovering objects and their location in images. In *ICCV*, 2005.
12. M. Villamizar, A. Garrell, A. Sanfeliu, and F. Moreno-Noguer. Online human-assisted learning using random ferns. In *ICPR*, 2012.
13. M. Villamizar, H. Grabner, J. Andrade-Cetto, A. Sanfeliu, L. Van Gool, and F. Moreno-Noguer. Efficient 3D object detection using multiple pose-specific classifiers. In *BMVC*, 2011.
14. M. Villamizar, A. Sanfeliu, and J. Andrade-Cetto. Orientation invariant features for multiclass object recognition. In *CIARP*, 2006.
15. M. Villamizar, A. Sanfeliu, and F. Moreno-Noguer. Fast online learning and detection of natural landmarks for autonomous aerial robots. In *ICRA*, 2014.
16. Michael Villamizar, Juan Andrade-Cetto, Alberto Sanfeliu, and Francesc Moreno-Noguer. Bootstrapping boosted random ferns for discriminative and efficient object classification. *Pattern Recognition*, 45(9):3141–3153, 2012.