Proc. Int. Workshop on Learning for Adaptable Visual Systems LAVS'04
St Catharine's College, Cambridge, UK, August 2004

Improved Object Categorization by Unsupervised Object Localization

Gerald Schweighofer Andreas Opelt Axel Pinz
Institute of Electrical Measurement and Measurement Signal Processing
Graz University of Technology, Schiessstattg.14B, A-8010 Graz, Austria

Abstract

Many approaches in object categorization require prior knowledge about the objects scale and location in the image and impose a lot of constraints on the used images. Also the probability of learning relevant data depends on the number and variety of training images. Furthermore, it is not ensured that just relevant data is learned that directly corresponds with the object category without any manual assistance. We present a new idea that utilizes unsupervised object localization based on a structure-from-motion approach. Using spatio-temporal information taken from image sequences allows a separation of relevant data. This leads to a high probability of learning relevant data with a distinct reduction of the computational complexity needed to obtain a final classifier. Results show that learning just relevant object data improves the object representation.

1. Introduction

Most of the current object categorization approaches ([1], [7], [6], [8], [16]) are somehow constrained, either by prohibiting highly cluttered background in the training images or within the allowed object positioning. Usually this is solved by manually preselecting the object or by assuming that the object is always prominently located in the training images. Opelt et al. [13] got good classification results without these constraints. Even so, the probability of learning relevant information depends on the number of training images. The experiments with the approach used in [13] showed that some percentage of the learned data is not directly interrelated with the learned object but at the most within some contextual relation to the object. The amount of the learned data that is not directly related with the object increases if all the positive training images contain nearly similar background whereas the negative training images show different backgrounds. The functionality of the learning model in [13] cannot cope with these training sets.

We present an approach that satisfies the task of learning a classifier ensuring that just relevant data is used even if the training images are disadvantageously chosen. This solution works on highly cluttered images within the task of generic object categorization. Additional we perform an adequate experimental evaluation, showing the benefits of our idea.

To ensure that just relevant data is used we need to segment the object from the background, which is impossible from any single 2D image. Therefore, our approach is based on videos of scenes containing the objects, taken by a moving camera. Based on structure-from-motion, we segment the individual frames by using spatio-temporal information to generate the 3D structure. Many methods exits that are capable to extract the structure of a scene ([2],[4],[5]).

We use a two step approach which first extracts the camera pose from an artificial target (for detail see [14]) for each frame. Then the structure is estimated by triangulation of tracked image features (see figure 1). From this reconstruction we cluster the features in 3D-space into distinct objects, which can be learned by the framework presented in [13].

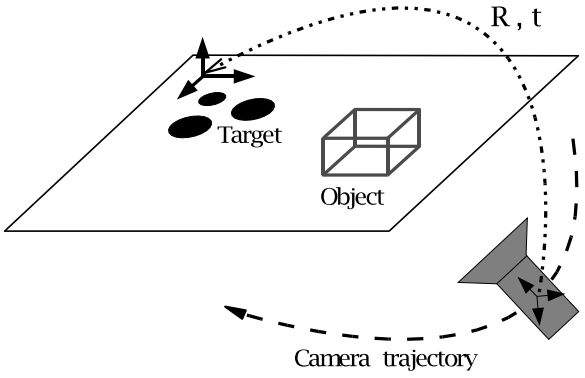
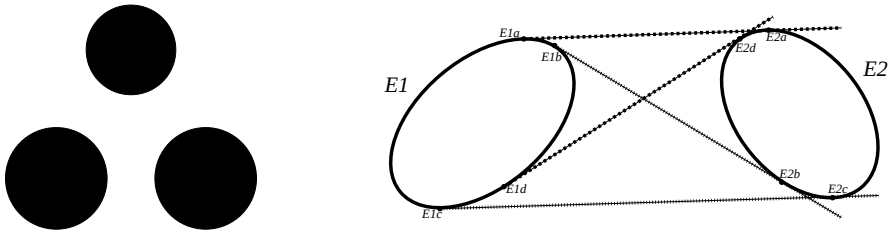


Figure 1. System setup showing the camera and the artificial target for camera pose estimation with respect to the object.



(a) (b)

Figure 2. (a) shows one of our targets. (b) shows that four tangents between two ellipses give eight invariant points under perspective projection.

2. Motion Estimation

Motion estimation is based on permanent camera pose updates, which estimate the 6 degrees of freedom of the camera (position t and orientation R) for each frame. The camera pose is estimated relative to the world coordinate frame, which is attached to our artificial target (see figure 1).

To compute the camera pose we need the 3D coordinates from a minimum of 4 coplanar image points. For this purpose, we have designed unique targets (see figure 2(a)) which are positioned in the scene. Under perspective projection, the invariant properties of such targets are the intersection points of the tangents between two circles (see figure 2(b)). A similar idea based on invariant projection of two circles has been presented by Chen et al. [3]. As we know the 3D coordinates of these points from our model, which also defines the coordinate system, we are able to estimate the camera pose. For computation we use an algorithm which is iteratively minimizing an error metric based on collinearity in the model. For more detail the reader is referred to [11].

3. Structure Estimation

To estimate the structure of the scene we use 2D-frame-per-frame correspondence. Natural image features (Harris corners [10]) are tracked over several frames. A point X_i in the scene is projected in the image j, with pose P_j, as image point p_{ij}

$$\lambda p_{ij} = P_j X_i. \tag{1}$$

As we know the pose P_j for each image from our target and the position of the point in more than one image we can calculate X_i. To eliminate the scale factor λ the cross product is calculated

$$p_{ij} \times (P_j X_i) = 0. \tag{2}$$

For each measured image point we obtain two independend equations from equation 2 in terms of X_i. To solve the linear system of equations we use singular value decomposition.

To take care of outliers we backproject the reconstructed point X_i in all images j and calculate the error as the distance between measured and reprojected points. If the error is larger than 10 times the median of all errors then the point is rejected as an outlier.

4. Object Localization

After estimating the 3D position of the corner features we try to combine adjacent 3D features to clusters which represent individual objects. As we do not know the number of visible objects in one video, the number of clusters is also unknown. Thus several clustering algorithms like k-means cannot be used. To cope with that problem we use a hierarchical clustering approach. A dissimilarity matrix for all reconstructed points X_i is calculated. The dissimilarity matrix contains the Euclidian distances from each point to all other points.

This matrix is used to iteratively connect close pairs of points (or clusters) to bigger clusters. Finally this process would result in one big cluster containing all points. Therefore, we need to define a stop criterion. We use the furthest distance between two points in a cluster. Two clusters will be merged if and only if the furthest distance of the merged cluster will be smaller than a certain threshold. In our experiments we set this threshold to 0.5 meters.

As a result of this clustering process we obtain a few clusters. In figure 3 we have backprojected all reconstructed 3D features in one image of a test video image sequence. The clusters are overlapping in the image. This happens only because of the projection into the image. As a difference to [15] we are only interested in one object, which is the one we want to learn. We recorded the videos only for this purpose with the constraint that the object (and also its cluster) is visible in all frames.

(a) (b)

Figure 3. The backprojected clusters in the image are shown in (a) whereas (b) shows just the found clusters.

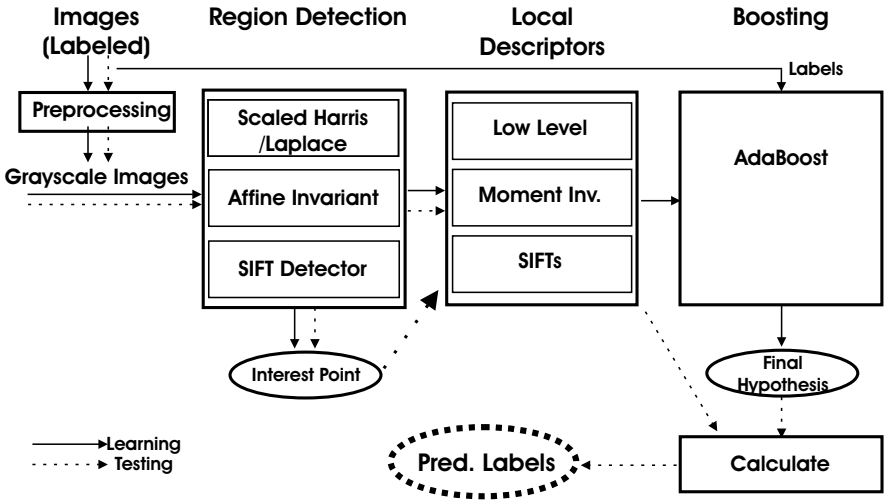


Figure 4. The framework used for the classification (for details see [13]).

5. Classification

To generate a classifier we use the framework presented in [13]. Figure 4 shows an overview of the framework. The labelled images are input to the region extraction module, which extracts regions of interest around salient points (features). These regions are normalized with respect to scale, affine transformations and illumination. The next module offers the possibility of calculating various description vectors for these regions. The images now represented by description vectors of salient regions are input to the learning module. The learning procedure based on AdaBoost computes a final classifier that is a linear combination of various weak classifiers that are obtained by the Weak Hypothesis Finder. The final classifier is then applied to new test images.

6. Experiments

In our experiments we used an affine invariant interest point detector ([12]). The resulting regions are normalized to a uniform size of 16x16 pixels. To describe the resulting regions we use moment invariants [9] with a dimension of 9.

As a dataset we created image sequences of 100 different books with nearly the same background (see figure 5).

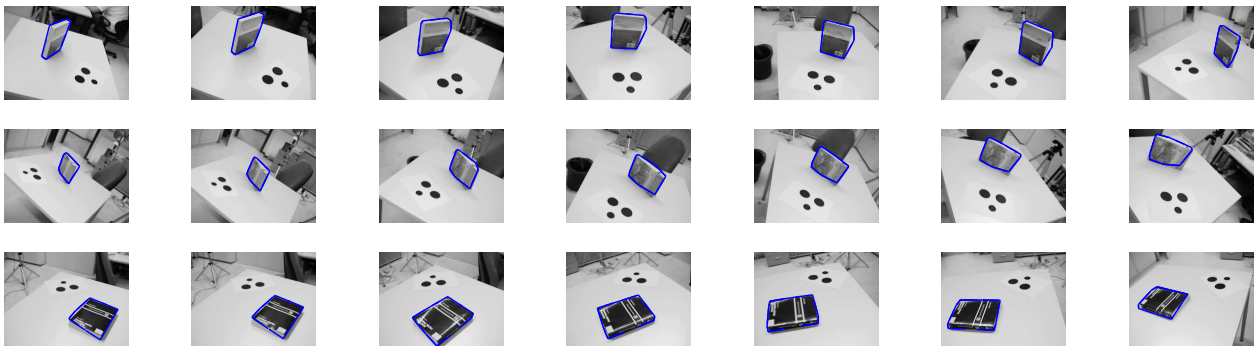


Figure 5. This figure shows examples from our dataset. Each row shows some frames of one image sequence of an instance of the category books.

Figure 6. This figure shows some examples from our background set.

As a negative class we used 300 images taken in our office (see figure 6), not containing the object and with different backgrounds. The results were obtained using 50 videos for training and 50 for testing, choosing the first, middle and last frame of each image sequence. The features located on the target were automatically removed. 150 images from the background set were used for training and the other half as test images.

We wish to compare the classifiers we achieve by training on the whole images against classifiers we achieve by only using interest points which lie inside the clusters.

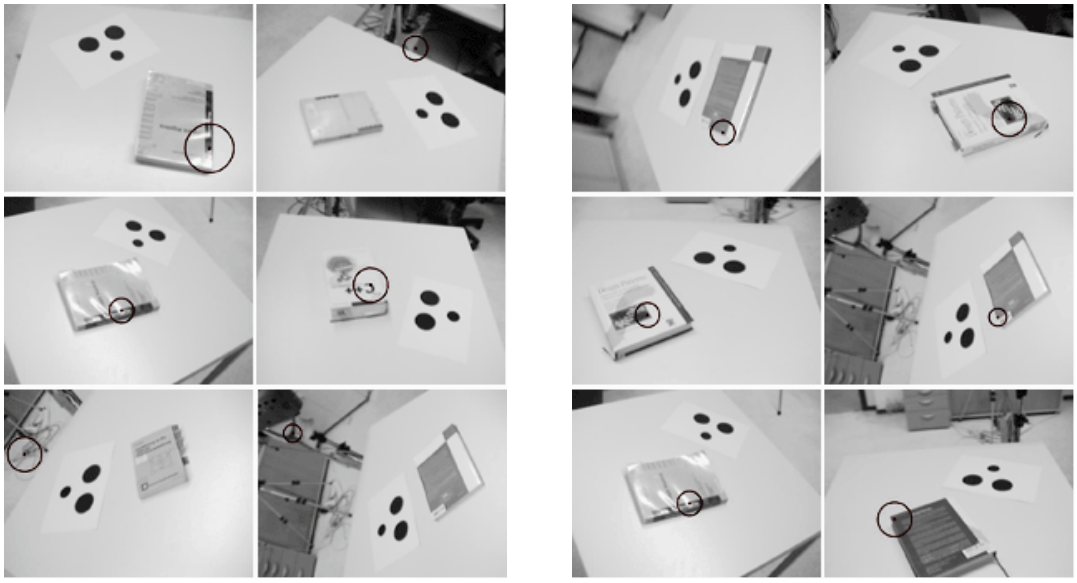
It is obvious that the classifier resulting from the training on the unclustered images might also contain irrelevant data. Choosing an average background distribution as shown in [13], approximately 20 percent of the learned weak hypotheses are not located on the object. But most of them are contextually related with the object. Problems occur if training is performed on an unfavorable dataset as in our case, where the backgrounds of the positive images are often similar but different to the negative training images. In our experiment the resulting classifier, that consists of several weak hypotheses contains more than 40 percent of irrelevant data. The first six weak hypotheses are shown in figure 7(a). In contrary to that we learned on the clustered images and obtained a classifier containing just relevant data. Figure 7(b) shows the first six weak hypotheses of this classifier where all the subsequent hypotheses are also located on the object.

With respect to the classification performance we created the ROC curves using the classifier learned on the unclustered images. Figure 8(a) shows a classification result of 86.3% testing on the whole test images. The solid line represents the ROC curve applying this classifier to the clusters in the positive images. The poor performance of nearly guessing shows that a classifier consisting of 40 percent irrelevant data might give good results but is not classifying the object but rather the background.

Figure 8(b) shows the ROC curves using the classifier learned on the clustered images. Again the dashed line represents the classification performance on the unclustered test images which increases to 87.5%. Using 100 percent relevant data in this classifier we obtain a result of 84.3% tested just on the object clusters in the images. The small difference in the performance is due to some background regions that are closely similar to some parts of books.

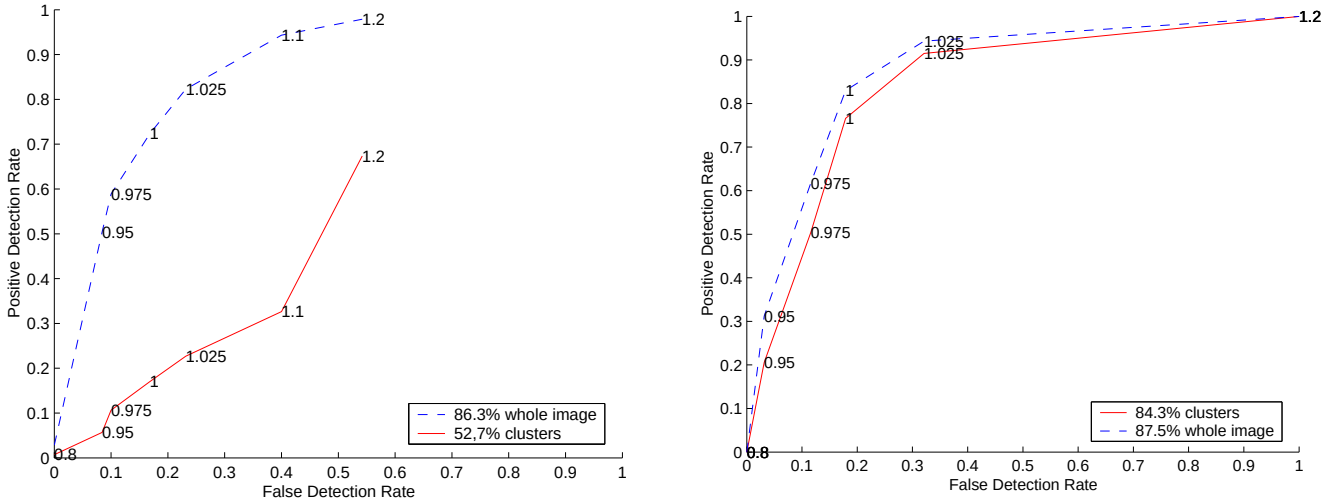
In table 1 we show a summary of the classification performances obtained using the various combinations of learning and testing.

Our Learning Module using AdaBoost results in a classification output of 1 or -1 voting for a belonging to the category or not, respectively. This result is obtained by applying a sigmoid function to the score produced by a linear



(a) (b)

Figure 7. (a) shows the first six weak hypotheses learned from the raw image data whereas (b) shows the first six weak hypotheses learned with the clustered image data.



(a) (b)

Figure 8. ROC curves for the classification on the dataset books. (a) shows the test result with a classifier learned with features in the whole images whereas (b) shows the results we obtain using just the clustered features.

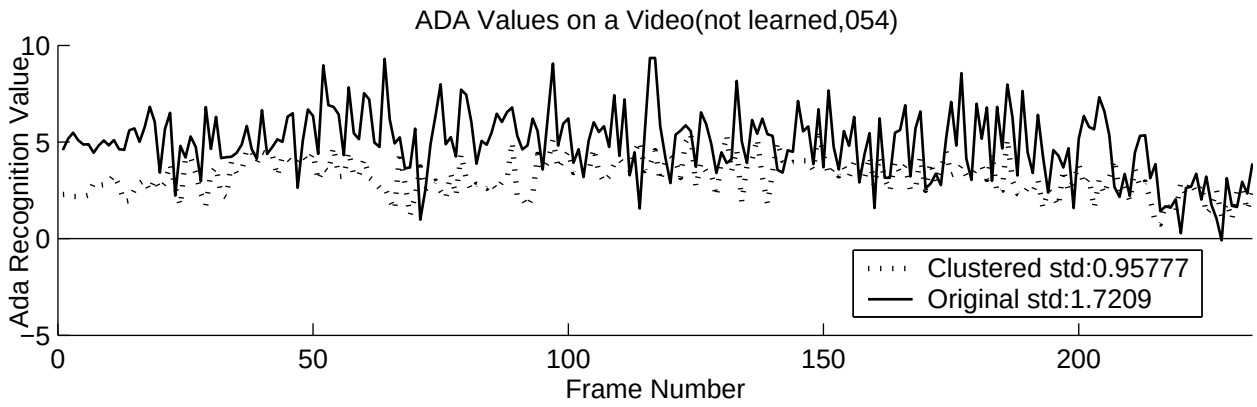


Figure 9. This diagram shows the resulting classification scores of AdaBoost for each frame of one video of the test set using the whole image. The solid line shows the variation of the score of the classifier trained on the whole image whereas the dashed line represents the scores tested with the classifier obtained on the clustered training images.

	Learning	
	whole images	image clusters
test on whole image	86.3%	87.5%
test on image clusters	52.7%	84.3%

Table 1. The classification results using the object clusters and the whole image for learning and testing.

combination of all the weak classifiers of this final classifier. Figure 9 shows the variation of this score over a whole image sequence of a book from the test set. This diagram shows that a classifier obtained by training on the whole image has a higher uncertainty in the classification result. Using a classifier trained on the clustered images for categorizing each frame of this video results in scores having half of the standard deviation than using the classifier trained on the whole image. Even if the values in the first row of table 1 are nearly the same the high standard deviation of the score means that the background has a high influence on the classifier trained on the whole images.

7. Discussion and Outlook

Object categorization using AdaBoost without any prior knowledge in the object location can result in uncertain classifiers if the training images are disadvantageously chosen. To cope with this problem we used a 3D reconstruction approach based on structure-from-motion on videos, combined with our object recognition framework. The object localization is used to cluster the region around the object. This enables the learning of just object relevant data to create a final classifier. The experimental evaluation shows that this suppresses the learning of classifiers that contain background data which leads to a categorization with high certainty of being directly related to the object.

This work is a first step towards combining spatio-temporal information with object recognition. We further want to do more experimental evaluation as well as learning object categories using objects that are represented through features that are stable through many frames of the video.

Acknowledgements

This work was supported by the Austrian Science Foundation (FWF, project S9103-N04) and by the European project LAVA (IST-2001-34405).

References

[1] S. Agarwal and D. Roth. Learning a sparse representation for object detection. In *Proc. ECCV*, pages 113–130, 2002.

[2] T. Brodský, C. Fermüller, and Y. Aloimonos. Structure from motion: Beyond the epipolar constraint. Technical Report CS-TR-4000, Computer Vision Laboratory, Center for Automation Research, University of Maryland, 1999.

[3] Q. Chen, H. Wu, and T. Wada. Camera calibration with two arbitrary coplanar circles. In *Proc. ECCV*, pages 521–532, 2004.

[4] K. Cornelis, M. Pollefeys, M. Vergauwen, and L. V. Gool. Augmented reality from uncalibrated video sequences. In M. Pollefeys, L. V. Gool, A. Zisserman, and A. Fitzgibbon, editors, *3D Structure from Images - SMILE 2000*, volume

2018 of *Lecture Notes in Computer Science*, pages 144–160. Springer-Verlag, 2001.

[5] U. Dhond and J. Aggarwal. Structure from stereo-a review. *IEEE Transactions on Systems, Man and Cybernetics*, 19(6):1489–1510, Nov.-Dec. 1989.

[6] G. Dorko and C. Schmid. Selection of scale-invariant parts for object class recognition. In *Proc. International Conference on Computer Vision*, 2003.

[7] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proc. CVPR*, 2003.

[8] R. Fergus, P. Perona, and A. Zisserman. A visual category filter for google images. In *Proc. European Conference of Computer Vision*, pages 242–256, 2004.

[9] L. V. Gool, T. Moons, and D. Ungureanu. Affine / photometric invariants for planar intensity patterns. In *Proc. ECCV*, pages 642 – 651, 1996.

[10] C. Harris and M. Stephens. A combined corner and edge detector. In *Proc. of the 4th ALVEY vision conference*, pages 147–151, 1988.

[11] C. Lu, G. Hager, and E. Mjolsness. Fast and globally convergent pose estimation from video images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(6):610–622, June 2000.

[12] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *Proc. ECCV*, pages 128–142, 2002.

[13] A. Opelt, M. Fussenegger, A. Pinz, and P. Auer. Weak hypotheses and boosting for generic object detection and recognition. In *European Conference on Computer Vision*, volume 3022, pages 71–84, 2004.

[14] G. Schweighofer, M. Ribo, and A. Pinz. Sparse 3d reconstruction of a room. In *Proc. of 28th ÖAGM / AAPR Workshop*, 2004.

[15] C. Stock, M. Lambrecht, A. Opelt, and A. Pinz. Object-centered feature selection for weakly-unsupervised. In *Proc. of 28th ÖAGM / AAPR Workshop*, 2004.

[16] J. Thureson and S. Carlsson. Appearance based qualitative image description for object class recognition. In *Proc. European Conference of Computer Vision*, pages 518–529, 2004.