A method for visual model learning during tracking

Miranda Grahl, Frank Joublin, Franz Kummert

2010

A method for visual model learning during tracking

Miranda Grahl[1], Frank Joublin[2], Franz Kummert[1]

[1]Cor-Lab, Bielefeld University, D-33615 Bielefeld/Germany
mgrahl,franz@cor-lab.uni-bielefeld.de
http://www.cor-lab.uni-bielefeld.de
[2]Honda Research Institute Europe GmbH, D-63073 Offenbach/Germany
Frank.Joublin@honda-ri.de
http://www.honda-ri.de

Abstract. In this paper, we propose a new method for visual model learning. The algorithm learns an object representation by one-shot and adaptively extends a set of saliency filters. The filter coefficients are extracted from the environment by different views. In addition, the algorithm fuses already learned visual filters and derives new visual classifiers in order to gain generalized object concepts. We evaluate our method on tracked sequences that are resulted from a processing with a visual bottom-up attention model.

Key words: adaptive learning, active vision, visual filtering

1 Introduction

For object learning with robots, it is necessary that a robot actively learns to extract visual filters from the environment. A visual filtering process in a dynamic scene requires the integration of multiple views with respect to environmental changes. In object recognition this is also known as object constancy. This motivates the investigation into incrementally object categorization during the tracking of a scene with focus on view-based object recognition. Visual classifiers that result from manually annotated training samples are not modifiable during the recognition process. This is reasoned by the fact that the object recognition system does not learn to structure the perceptual information by itself. In order to avoid these drawbacks, a system needs to extract and to learn those view-based classifiers unsupervised from its environment. Less work has been investigated into object learning from active vision perspective with respect to an adaptation of an appropriate active visual filtering process. Moosmann et al. [1] propose a learning of a visual filter based on a biased decision tree. A further object classifier is learned from few training samples in [2] on the basis of a Bayesian framework. Both methods categorize an object without focusing on its central region and miss the incremental integration of object views during tracking into one classifier. Walther et al. [3] define a selection of salient regions as a reliable basis for object recognition and determine an object representation

by a set of SIFT features [4]. A complete bottom-up attention mechanism would fail for an object representation during observing the scene. The integration of continuous changing information is missing e.g. the different views of a rotating hand. Schendan et al. [5] report that a categorization of unusual views of objects requires more reaction time. This is also reported by Ganis et al. [6] who suggest the integration of top-down processes in the categorization process.
In this paper, we present a new method for visual model learning during observing a scene. We assume a minimum requirement for object learning and bootstrap the learning process by a visual bottom-up attention model [7]. The method extracts additional visual classifiers and hypothesizes object constancy in the center of the observed scene. This center information is used as a supervised signal during the tracking in order to approve the validity of an extracted classifier. Our approach obtains an object representation by a linear combination of possible stored views. After a new observation, the method enables a fusion of similar responding stored models and retrieves them for the filtering process.
The paper is structured as follows. In section 2, we present the learning method for the visual filtering process. Section 3 focuses on the evaluation with respect to different object views. We give a conclusion in section 4.

2 Learning Method

The default visual aspects are defined by a set of filters [7] that extract a salient point that is kept in the central region of the current view during the track. A new *saccade* is triggered by a timer event and the gaze is recentered on a salient point. Nonlinear view based models are extracted by SIFT features. During a saccade, a visual model f_j is learned by an one-shot learning process from the centered image $I(x, y)$. The one-shot learning process defines a filter as a triple $f_j = \{s_{ji}, w_{ji}, c_j\}$ for the *j-th* filter that is learned during the tracking. It consists of a set of SIFT features s_{ji}, learned weights w_{ji} and a learned weighting coefficient c_j for this filter. Each tracking step is evaluated by gaining an improved presentation of a visual classifier by hypothesizing object constancy. During the tracking the object hypothesis is approved and an integration of new visual filters is realized. During a saccade the activity of already learned saliency filters are compared and fused in order to achieve a generalized classifier.

2.1 One-shot learning of a visual filter model

In a first step a set of SIFT features is selected (see fig. 1). The *SIFT extraction* defines each position of I as a keypoint in order to extract a sift vector s with a defined σ. Each centered image results into a set of overlapping SIFT features S that describe the local orientation characteristics. The center feature and peripheral features for the *j-th* visual filter in the current track are defined by s_{j0} and s_{ji} with $i = 1 \ldots n_j$. Each s describes a nonlinear view-based model. After the extraction of S, the *one-shot learning* process of an object view is initialized

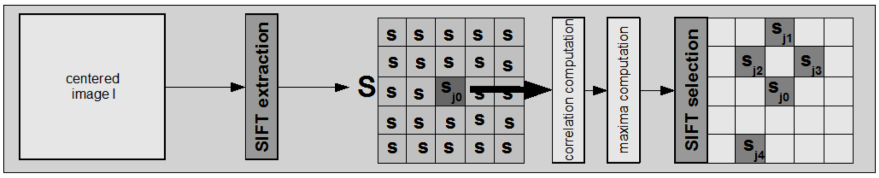

Fig. 1. Selection of nonlinear view-based models for the one-shot learning with an inhibition.

by the selection of nonlinear view-based models s_{ji}. The idea of the *SIFT selection* bases on the assumption to enhance the specificity of an extracted filter. Therefore, those locations are inhibited in the periphery that correlates strongly with s_{j0}. For this a *correlation* (1) with s_{j0} is computed yielding a feature map $\varPhi_{j0}$ that is subject to a local *maxima search.*

$$\varPhi_{j0} = \sum s_{j0} \cdot S \; . \tag{1}$$

Those s_{ji} $\left(i = 1, \ldots, n_j^{sel}\right)$ that exhibit a large correlation value are extracted and processed (for one-shot learning see fig. 2). This selection step results into a set of nonlinear view based models that contain one positive s_{j0} for the center position and a set of negative s_{ji} for the inhibition. After the selection of s_{ji} a *correlation computation* with S results into a set of feature maps $\varPhi_{ji}$ (2).

$$\varPhi_{ji} = \sum s_{ji} \cdot S \; . \tag{2}$$

These features maps are used for the weight initialization for the visual model.

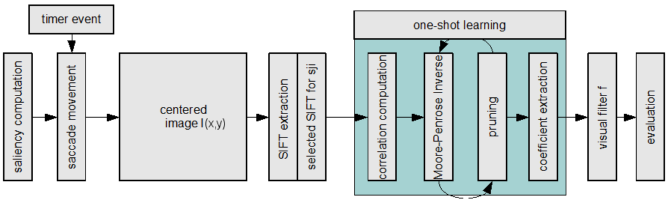

Fig. 2. Visual filter learning process during tracking.

The weights w_{ji} are computed by the *Moore-Penrose Inverse* + of $\varPhi_{ji}$:

$$w_{ji} = \varPhi_{ji}^{+} \, g \text{ with } g = exp\left(\left(-x^2 - y^2\right)/\sigma^2\right) \; . \tag{3}$$

Afterwards an additional *pruning step* is conducted for those weights w_{ji} with positive values. The step removes false positive w_{ji} and corresponding $\varPhi_{ji}$ and recomputes w_{ji} again.The initial response of f_j is determined by a saliency map

y_j (4) that is defined by the linear combination of w_{ji} and Φ_{ji}.

$$y_j = \frac{1}{c_j} \cdot \sum_{i=0}^{n_j^{sel}} w_{ji}^T \, \Phi_{ji} \ . \tag{4}$$

The weighting coefficient c_j is derived by a weighted mean (5) from the expected center activity a_j in the initial phase.

$$c_j = \sum_{x,y} g \cdot \sum_{i=0}^{n_j^{sel}} w_{ji}^T \, \Phi_{ji} \text{ and } a_j \ = \sum_{x,y} y_j \cdot g \ = \ 1 \ . \tag{5}$$

During the tracking the centered image is convolved with f_j according to equation (4) and the resulting center activity $a_j^{'}$ is compared to a_j. A new filter is inserted, if the current y_j does not fulfill the object constancy hypothesis i.e. $a_j^{'} < \theta_1$. This means during the observation of an object the method integrates several subviews f_{j^*} into one visual classifier $\mathbb{F}_k$ related to y_{j^*} (6).

$$y_{j^*} = \max_j (y_j) \ \text{ and } a_{j^*} = \sum_{x,y} g \cdot y_{j^*} \ . \tag{6}$$

2.2 Fusion and initialization of new visual filters

During a saccade an *evaluation* step is conducted that decides whether a new filter should be added or already learned filters should be used. Besides the integration of different views of an object during the tracking, a fusion combines learned filters with two strategies. The first strategy approves the center activity $a_{j^*}^{'}$ and decides if an already learned classifier $\mathbb{F}_k$ for object k is used again for the filtering process for the current observed object with $\mathbb{F}_k = \mathbb{F}_k \cup \ \{f_{j^*}\}$. If more than one visual filter $\mathbb{F}$ responses above a defined threshold θ_2, they are combined into one classifier $\mathbb{F}_k = \mathbb{F}_k \cup \mathbb{F}_l$. In the case that no classifier is available for the current observed object ($a_{j^*}^{'} < \theta_2$), a new visual classifier is defined by $F_{k+1} = \{f_{j^*}\}$.

3 Evaluation

The evaluation of our method bases on a video sequence that shows a person who demonstrates a cup stacking task. A saccade movement is determined from a saliency map that is computed by color, orientation, motion and intensity. An inhibition of return leads to a gaze selection that have not been attended before. A new saccade is triggered each second and separates the demonstrated task in tracked sequences. Those tracked sequences are removed from the dataset with more than one object in the center. A learning from multiple objects will be one aspect of the further development of the proposed algorithm. For the extraction of SIFT features, we resize the images from 525x525 pixels to 159x159 pixels and compute them with $\sigma = 2$. A SIFT vector describes each pixel in 8 orientations for 4x4 spatial bins. The gaussian kernel g is computed with $\sigma = 0.05$.

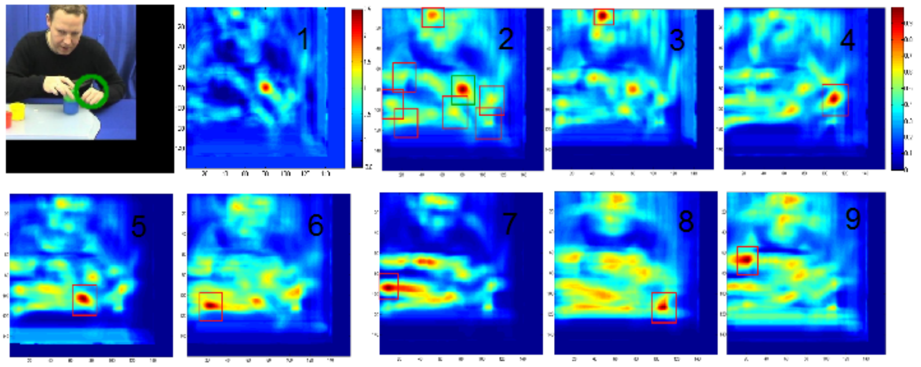

Fig. 3. Example of the one-shot learning method (from left to right). The image shows a hand (green circle) in the current centered observation. The image 1 shows the saliency map y_j of the one shot learned model with $w = (0.57, -0.08, -0.08, -0.11, -0.07, -0.02, -0.03, -0.10)$, a_j=1 and c_j=0.11. The image 2 shows the maxima extraction and feature map θ_{s0} (see eqn. 1). The green square shows s_{j0}. Red squares mark s_{ji} for the inhibition and corresponding θji (3-9) (see eqn. 2) are shown.

3.1 One shot learning method and insertion of subviews

At first we evaluate the performance of a learned one-shot visual model f_j and the insertion of subviews j into one classifier $\mathbb{F}_k$ during a tracking step. The performance of both is compared to the performance of a simple visual model f_{j0} without an inhibition of peripheral observations with s_{j0} and $w_{j0} = 1$. An example of the one-shot learning method is shown in figure 3. The comparison of image 1 and 2 shows that the learned visual model inhibits regions like the face and cups and enhances the specificity in the center field.
Figure 4 shows the integration of additional filters f_j with respect to changing views of a hand. The insertion is marked with enlarged pictures (from left to right). The phases 1 and 2 marks the center activity $a^{'}_{j*}$ before and after the insertion and corresponding saliency maps y_{j^*}. In a first step f_j is learned by one-shot and three additional filter models are gradually inserted into one classifier $\mathbb{F}$. Phase 1 shows that the current filter is no longer valid and $a^{'}_{j*}$ decreases. In phase 2 the center activity $a^{'}_{j*}$ is improved by the insertion of f_j. The modification of the filter model leads again to a specific response to the observed hand.

In order to evaluate the performance of our approach, we compare the different averaged center activities of f_{j0}, f_j and the resulting classifier $\mathbb{F}_k$. The visual filter f_{j0} and f_j are derived from the first image of the tracked sequence from the insertion process. The performance is tested on two datasets which contain *true positive tp* and *true negative tn* hand samples (see fig. 5) in the center view. The top right dataset contains 33 (*tp*) samples and the dataset bottom right contains 22 (*tn*) samples without the appearance of a hand. Both activities of

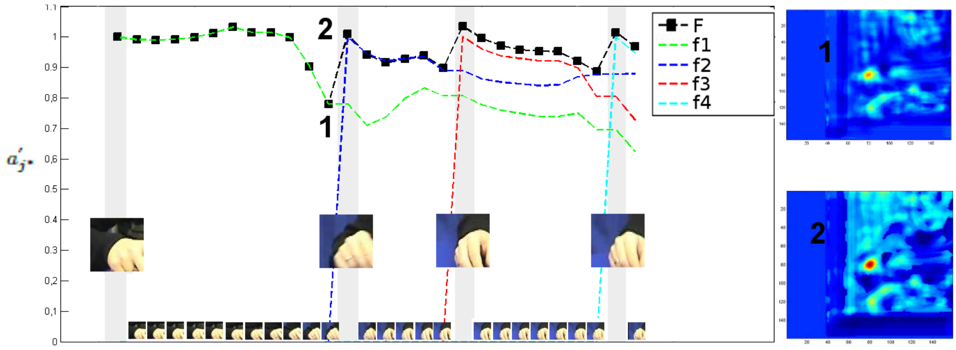

Fig. 4. Example of insertions (gray) of f_j during tracking into one visual classifier $\mathbb{F}_k$. Image 1 shows the corresponding y_j^* (below the threshold $\theta_1 = 0.9$). Image 2 shows corresponding y_j^* after insertion.

the simple model f_{j0} show a high value to both datasets. The generalization of detecting hands of this classifier is high. But also reveals a high activity with respect to the absence of a learned hand model. In contrast to this, f_j and $\mathbb{F}_k$ suppresses the activity for objects to other classes. Both show a high activity to different hand views, where $\mathbb{F}_k$ performs better than f_j.

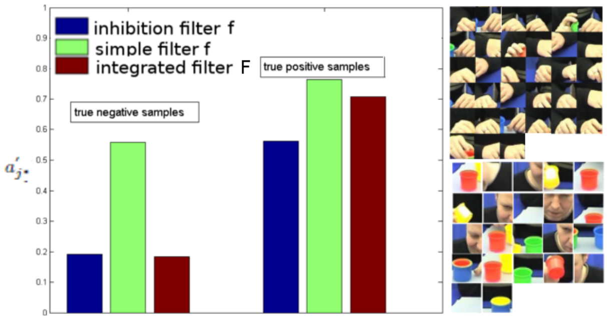

Fig. 5. Comparison of a simple visual model f_{j0} (green), f_j (blue) and $\mathbb{F}_k$ (red) with respect to averaged a_{j^*}. Images at top comprise samples with hands. Images below comprise samples of different attended locations.

3.2 Fusion of visual models

In a first step, we show the fusion process of two visual classifiers $\mathbb{F}$. The evaluation bases on six tracked sequences that captures a left and a right hand. In a second step, we apply our learning method on tracked sequences that exhibit different objects. The fusion of visual classifiers is depicted in figure 6 (from

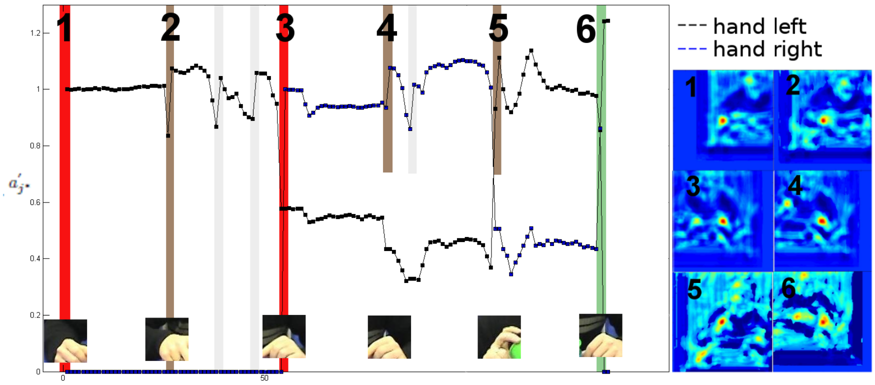

Fig. 6. Fusion of visual models during tracking and corresponding a_{j^*} with $\theta_2 = 0.75$. Different color bars show the decision phases of the filter learning process during a saccade. The red bar marks a complete new a visual filter, brown for the fusion of a new learned visual model with an already learned visual model, gray for the insertion, green marks the fusion of two already learned visual filters with a new learned one-shot model. On the right hand side the corresponding saliency maps y_j^* are shown.

left to right). The small image patches show always the starting position of the tracking. The different phases of the learning method are depicted and colored. The corresponding filter response y_{j^*} is shown on the right hand side. The black and blue line shows the activity course for the left and right hand. In phase 1-2 learned one-shot models are fused in an improved visual classifier for the left hand. In phase 3 a new filter for the right hand is extracted. In phase 4 and 5 already learned filters are fused again for a separate recognition of both hands. In phase 6, both filters for the left and right hand show a similar activity $a'_{j^*} > \theta_2$ and fuse into one visual classifier. The resulting saliency map is shown in image 6. The filter is now able to classify both hands. The filter learning during the tracking of different objects is depicted in figure 7 (from left to right). These objects are cups, hands and faces. We evaluate 15 sequences that result into 6 visual classifiers. The figure shows the activity a'_{j^*} of learned visual filters during a saccade and the corresponding location for gaze fixation. The learning process incrementally adds new visual classifiers, where the activities are shown with colored bars. The filters for the right and left hand slowly converge into one classifier. The responses of other filters show less activities in case of fixating hands. Cups and faces also develop visual filters and show less activities with respect to the absence of a learned visual model.

4 Conclusion

In this paper, we propose a new method for a visual filtering process that learns incrementally a view-based object representation. Based on a saliency computation, our approach assumes a minimum of requirements and learns a set of

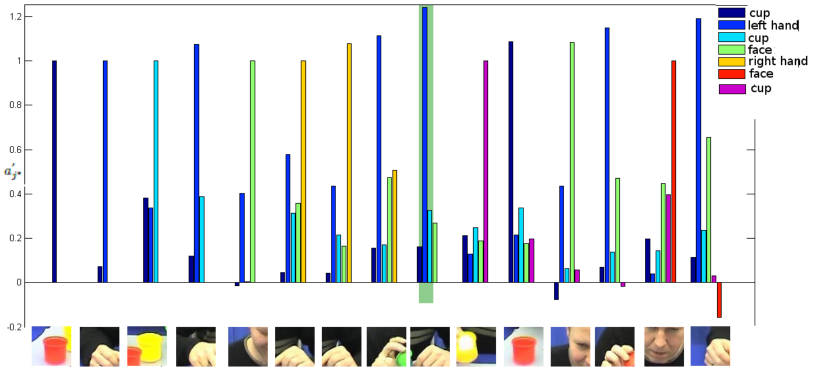

Fig. 7. Activity of different learned visual filters during scanning the scene with θ_1=0.9 and θ_2=0.75. After a fusion (green bar), the left hand filter responses to further observed objects and shows a hight activity to hands.

saliency filters. Our method shows accurate results with respect to the absence of a learned model. Secondly, our method improves a visual model during the tracking by insertions of different views. Thirdly, the proposed fusion method combines several classifiers and enables a recognition of two hands that are separately recognized in an initial learning phase.

Acknowledgments. The work described was supported by the Honda Research Institute Europe.

References

1. Moosmann, F., Larlus, D., Jurie, F.: Learning saliency maps for object categorization.In: European Conference in Computer Vision (ECCV), International Workshop on The Representation and Use of Prior Knowledge in Vision (2006)
2. L. Fei-Fei, R. Fergus, and P. Perona: A Bayesian approach to unsupervised one-shot learning of object categories. In: Proc. ICCV, 1134–1141 (2003)
3. Walther, D., Rutishauser, U., Koch, C., Perona, P.: Selective visual attention enables learning and recognition of multiple objects in cluttered scenes. In: Computer Vision and Image Understanding, 100(12), 41–63 (2005)
4. Lowe, D.: Object recognition from local scale-invariant features. In: Proceedings of the 7th International Conference on Computer Vision, 1150–1157 (1999)
5. Schendan, H.E., Stern, C.E.: Where vision meets memory: prefrontal-posterior networks for visual object constancy during categorization and recognition. In: Cereb. Cortex. 18, 1695–1711 (2008)
6. Ganis, G., Schendan, H. E., Kosslyn, S. M.: Neuroimaging evidence for object model verification theory: Role of prefrontal control in visual object categorization. In: Neuroimage, 34, 1384–1398 (2007)
7. Itti, L., Koch, C., Niebur E.: A model of saliency-based visual attention for rapid scene analysis. In: IEEE Transactions on Pattern Analysis and Machine Intelligence 20(11), 1254–1259 (1998)