4th International Scale Space Conference, Isle of Skye, UK, June 2003

Computation of generic features for object classification

Daniela Hall*, and James L. Crowley

GRAVIR–IMAG, INRIA Rhônes–Alpes
38330 – Montbonnot Saint Martin, France

Abstract. In this article we learn significant local appearance features for visual classes. Generic feature detectors are obtained by unsupervised learning using clustering. The resulting clusters, referred to as "classtons", identify the significant class characteristics from a small set of sample images. The classton channels mark these characteristics reliably using a probabilistic cluster representation. The classtons demonstrate good generalisation with respect to viewpoint changes and previously unseen objects. In all experiments, the classton channels of similar images have the same spatial relations. Learning of these relations allows to generate a classification model that combines the generalisation ability from the classtons and the discriminative power from the spatial relations.

Keywords: local image features, classifi cation, clustering

1 Introduction

Structural matching is a classical approach for object recognition. Gaussian derivatives measure the basic geometries of the appearance of local features. In such a feature space, similarity of features can be measured by the distance between their vectorial representation. This feature matching principle is widely used for image indexing, and object identifi cation [8,14].

Classifi cation is a task that requires the assignment of previously unseen objects to the corresponding class of visually similar objects. Classical feature matching fails in many cases due to large feature variations among objects a class. For this reason vision systems have diffi culties to generalize from a small set of images to other images of the same class.This makes classifi cation a much harder problem than identifi cation of previously seen objects.

Successful classifi cation relies on the extraction of signifi cant class features that should be robust to changes in viewpoint, object identity, position, scale and lighting conditions. This article adresses the problem of the extraction of such signifi cant features. Generic feature detectors have the property that they mark the most characteristic features with respect to a learned class. In our method, the generic features are computed automatically by unsupervised clustering. We propose a measure for the selection of the most signifi cant clusters and several experiments show that the selected clusters detect those signifi cant features robust to changes in viewpoint and object identity.

* This research is funded by IST CAVIAR 2001 37540

2 Composition of generic features (classtons)

The idea of vector quantization or clustering of the outputs of linear filter sets has been applied by Leung and Malik for texture recognition and image segmentation [6,9]. They define texture as entity with spatially repeating properties. Zhu and his collaborators obtain clusters robust to rotation and scale changes by applying a transform component analysis to image patches before clustering [15]. The obtained textons that represent the texture clusters allow the efficient modeling of textures. Schmid has applied the same k-means clustering scheme to compose generic features for image indexing [13]. We want to extend this idea and use exclusively clusters in feature space for image description, recognition and classification.

A visual object class consists of visually similar images with spatially repeating properties over these images. Under these constraints the clustering of vector representations of local features is able to detect automatically the repeating features and learn their variations. Clustering is therefore a means for the computation of the desired generic features.

3 Clustering approaches

The success of classification depends on the generic features (the classton vocabulary). The choice of an appropriate clustering algorithm is crucial. In this section we evaluate k-means, k-means with pruning and DBScan. The methods are compared on several test databases.

The choice of the comparison of those three methods is motivated by the work of Leung, Malik, Schmid, and Zhu, who all use k-means. Leung [6] uses k-means with pruning. This method is less sensitive to cluster center shifts due to outliers than the original k-means algorithm. We compare these standard methods to a new clustering algorithm from the data mining community. Ester [1,2] developed DBScan for the expansion of density clusters of arbitrary shape with a minimum of domain knowledge. The definition of DBScan allows to find natural boundaries between clusters. This property has the effect that the number and the shape of the significant feature clusters is automatically adapted to the data.

3.1 K-Means clustering

K-means is an agglomerative clustering method with a specific objective function. Assuming that there are k clusters and each cluster is represented by its center of gravity, an objective function is obtained by evaluating the distances of image points, $x_j \in D$, to their respective cluster center, c_i:

$$E(C,D) = \sum_{i=1}^{k} \left(\sum_{j \in C_i} (x_j - c_i)^T (x_j - c_i) \right), C = \{C_1, \ldots, C_k\} \qquad (1)$$

The algorithm assigns each point to the closest cluster center and updates the centers. These steps are iterated until the objective function reaches a minimum. K-means

has linear complexity $O(n)$ with n number of points. The simplicity and the effi ciency of the algorithm explains its popularity.

In most applications the optimal number of clusters is unknown. The objective function is proven to converge to a local minimum, not the global minimum. The problem of convergence to a suboptimal solution can be overcome by running k-means many times and retaining the best solution. This multiplies the computation time and there is no guarantee of the quality of the solution [3].

Standard k-means has the disadvantage that data points are assigned to the currently closest cluster center. Outliers as a result to noise are always present in the data. In extreme cases, the assignment of outliers shifts signifi cantly the center of gravity of the cluster and decrease the overall quality of the solution. K-means with pruning is less sensitive to outliers.

3.2 K-Means with pruning

In the fi rst step, standard k-means is applied with a large number of clusters (in the range of 500 to 8000 clusters). Then these clusters are reduced subsequently. Close clusters are merged and clusters with few elements are suppressed. This pruning step takes as parameters the distance ε between clusters for merging and a number of required elements $MinPts$. The pruning step is repeated with increasing ε until the desired number of clusters is reached.

This algorithm is computationally more expensive, because the data is represented by a larger number of initial clusters requiring more iterations. Subsequent merging and discarding of small clusters allow to assign fewer outliers to the clusters. The remaining clusters are more representative for the image characteristics.

3.3 DBScan clustering

In this section we describe an alternative clustering algorithm that is based on expanding clusters from a seed. This algorithm, referred to as DBScan, has been proposed by Ester [1,2] for the organisation of spatial databases with minimal requirements of domain knowledge. The algorithm is density based and can discover density clusters of arbitrary shape. This algorithm is interesting for the computation of classtons, because the number of clusters is determined automatically.

The key idea of DBScan is that the neighborhood of cluster points has to contain a minimum number of data points $MinPts$. In other words, the cardinality of a sphere with radius ε has to exceed the threshold $MinPts$. Such points can serve as seed for cluster expansion. The algorithm is formalised by following defi nitions.

Definition 1 (directly density reachable) A point p is directly density reachable from point q with respect to ε and $MinPts$ in the point set D if

- $p \in S_\varepsilon(q)$
- $card(S_\varepsilon(q)) > MinPts$

Definition 2 (density reachable) A point p is density reachable from q with respect to ε and $MinPts$ in D, denoted as $p >_D q$, if there is a chain $q = p_1, \ldots, p_n = p$ such that $p_i \in D$ and p_{i+1} is directly density reachable from p_i (see Figure 1).

Definition 3 (density connectivity) p is density connected to q with respect to ε and $MinPts$ in D if there is a point $t \in D$ such that both p and q are density reachable from t (see Figure 1).

$$card(S_\varepsilon(q)) <= MinPts \tag{2}$$

$$\exists p : q >_D p \text{ and } card(S_\varepsilon(p)) > MinPts \tag{3}$$

A cluster is defined as a set of density connected points which is maximal with respect to density reachability. Noise is the set of points that are not contained in any cluster. A cluster contains core and boundary points. Core points are those points that fulfill $card(S_\varepsilon(q)) > MinPts$. Boundary points are points that fulfill

To find a cluster, DBScan starts with an arbitrary point p and retrieves all density reachable points. If p is a core object, this yields a new cluster. If p is a non-core object, no points are density reachable from p and it is assigned to noise. In the first case, the density reachable points are used to expand the cluster until maximality. The algorithm continues until all points are labelled.

DBScan has the advantage that every point is treated only once, by computing $card(S_\varepsilon(p))$ and assigning the point to the current cluster or to noise. A sophisticated implementation of DBScan has a computational complexity of $O(n \log(n))$. The number of clusters is determined automatically. Standard k-means does not reject outliers which decreases the overall quality of the clustering. DBScan detects outliers automatically and assigns them to noise.

3.4 Evaluation

K-means with pruning has a computational complexity of $O(nk)$, with n number of points and k number of clusters. For standard k-means, the number of clusters is small which results in a linear complexity of $O(n)$. DBScan needs to compute the nearest neighbors for every point. In the current straight forward implementation the nearest

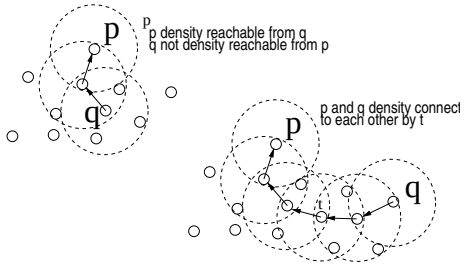

Fig. 1. Density reachability and density connectivity

neighbor search requires $O(n)$. By using for example binary search trees, the complexity can be reduced to $O(\log(n))$. The overall complexity of the current implementation is $O(n^2)$.

The computational difference becomes clear in following experiment. The results of the proposed clustering algorithms are evaluated on two data sets. Data set A has 63000 local image features extracted according to section 4 from four frontal face images from the AR face database [10]. Data set B consists of 9600 local image features extracted from four toy car pictures. The toy cars are segmented from the background, the faces images are unsegmented.

Table 1 displays processing times (on a 600MHz Pentium III). Standard k-means is fast, but the optimal number of clusters is unknown. To ensure that a good solution is found, k-means should be run several times with changing k. This multiplies the computation costs. K-means with pruning requires more iterations in order to minimise the overall error. Pruning is then called subsequently with parameters ε and $MinPts$ until the desired number of clusters is reached. For segmented images, DBScan is faster than k-means with pruning.

	k-means	k-means with pruning	DBScan
A (faces)	27 s, 6 iterations	1442 s, 27 iterations	7698 s, no iteration
B (toy cars, segmented)	12 s, 13 iterations	1712 s, 28 iterations	143 s, no iteration

Table 1. Processing time of image features from (A) unsegmented and (B) segmented images.

4 Feature prototype generation

In this section we describe the feature space used for feature extraction. In section 4.2 different cluster representations are evaluated. In order to select those clusters that correspond to the desired signfi cant features we need to be able to evaluate the quality of a particular cluster. Appropriate measures are proposed in section 4.3.

4.1 Feature description

Gaussian derivative receptive fi elds are used by many researchers for the description of local feature appearance [4,8,11,12,14]. Low order derivatives measure the basic geometries of features [5]. Local features are represented by the response to a bank of Gaussian derivative receptive fi elds centered on the image position. The receptive fi elds are scale invariant due to normalization for intrinsic scale. We compute the intrinsic scale as an extremum in the normalised Laplacian over scale as proposed by Lindeberg [7].

We experiment with following feature spaces: fi rst and second order derivatives and fi rst, second and third order derivatives. The suppression of the derivative of order zero makes the feature less sensitive to illumination variations. The features are extracted

either at a fixed scale, or at the specific intrinsic scale. In the second case, only those features are considered that actually display a maximum over scale within the predefined range (in our experiments $\sigma \in [1.34, 8.00]$). The data is normalised to compensate for the dynamic of receptive fields of different orders such that the distribution has 0 mean and 1.0 standard deviation.

Scale normalized features are clustered without taking into account their local scale. This has the advantage that features that occur at different scales due to perspective transformation are assigned to the same cluster. On the other hand, the relative scale between features of the same objects is lost. The scale relations between characteristic features of an object is discriminant and worth preserving. Instead of using the feature space $(L_x, L_y, L_{xx}, L_{xy}, L_{yy})$, we propose to use $(L_x, L_y, L_{xx}, L_{xy}, L_{yy}, \sigma)$. This adds the local scale to the feature space. A relatively scale invariant object representation is obtained that preserves the internal scale relations.

Figure 2 shows examples of clusters obtained from the toy car example using k-means. The linear combination of the cluster prototypes are shown. By comparing the left figures with the right figures, it can be observed that the classtons on the left and the classtons on the right display the same basic geometries. This means that the extension of the feature space to third order derivatives does not increase significantly the ability to describe the present geometries. A feature space up to second order covers sufficiently the geometry of the local features in the experiments.

4.2 Cluster representation

To enable classification we need to compute classton channels. In a classton channel those points that belong to the classton are marked. The generation of the classton channel requires an assignment algorithm that decides if a particular feature belongs to the classton. The quality of the assignment is closely related to the cluster representation.

The assignment can be computed by several algorithms. Many researchers use minimum distance to prototype [6,9,13]. This is a fast measure, but it does not take into account the distribution of the cluster points. This measure is acceptable when the clusters have close spherical shape or clusters are sufficiently far from each other. In the typical case where clusters are elliptic point clouds, a better representation is obtained by using a probabilistic measure based on the Mahalanobis distance.

$$p(C_i|x) = \exp\left(-(x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i)\right) \tag{4}$$

where Σ_i is the covariance matrix of cluster C_i. Clusters of arbitrary shape can be represented by a set of elliptic point clouds.

Figure 3 illustrates the effect of the different representation methods. The labels of the assigned cluster are coded as grey values. We observe connected regions in the channel images. This is due to the effect that spatially close features have similar appearance and are assigned to the same cluster. The connected regions of the probabilistic method are more stable than the regions obtained by distance to prototype. Although, in this example the differences are not significant, the probabilistic measure should be used to reduce incorrect assignments.

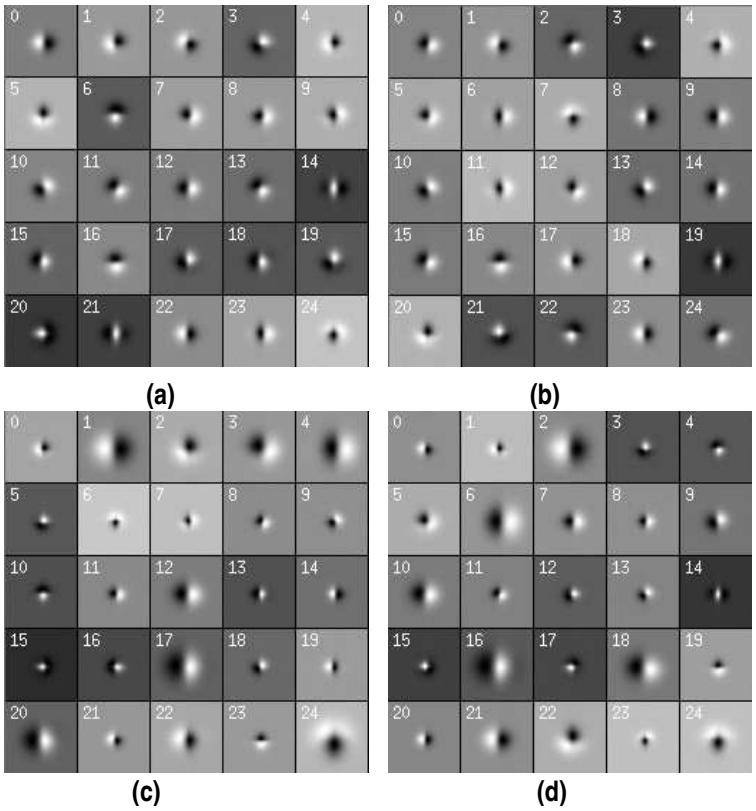

Fig. 2. Cluster examples. (a) feature space up to order 2. (b) feature space up to order 3. (c) feature space up to order 2 with scale. (d) feature space up to order 3 with scale.

4.3 Quality of clustering results

Clusters are by definition dense collections of data points. They are useful for classification because they represent a collection of highly similar features. Under the condition that the training images are visually similar, those dense clusters represent the most significant features for the trained image class.

Several parameters can be used to judge the quality of a cluster, such as the density and the compactness in feature space and the connectedness of the regions in the classton channel. The density of clusters depends on the total number of feature points. For this reason, a threshold for reliable detection of dense clusters can not be found. Compactness has the advantage that it is independent from the number of features, under the condition that the learned features represent sufficiently the true feature distribution. A generic feature with good generalisation ability produces large connected regions. Figure 5 shows an example of compact and connected clusters that specify forehead, hair, eyes, nose, and lips as significant features of faces.

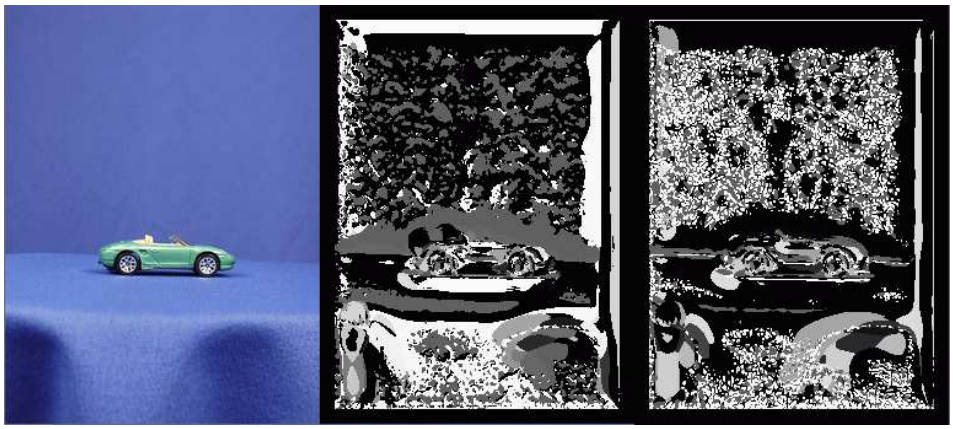

Fig. 3. Effect on feature assignment using different cluster representations (probabilistic or center of gravity).

Connectedness of regions can be measured as the average number of pixels per region. Compactness is defined as the ratio of a volume and the enclosing sphere. In order to compute the compactness of a point cloud, we modify the geometrical definition of compactness as follows:

$$\text{Compact}(C_k) = \frac{\prod_{i=1}^{N} \sigma_i}{\max_i(\sigma_i)^N} \tag{5}$$

The volume of a cluster C_k is approximated by the product of standard deviation of its members in each dimension $i = 1, \ldots, N$. The volume of the enclosing sphere is computed as the maximum standard deviation to the power of N. Density can be computed as average number of points per volume unit.

Connectedness allows to reduce the image to a number of regions. This gives the required tolerance to region positioning that enables classification robust to viewpoint changes and object identity. Figure 5 and Figure 6 show the classton channels of the most compact and the most connected classtons computed from 15 frontal faces and 13 toy cars from 2 viewpoints respectively.

5 Experiments and observations

5.1 The test database

We use three different test databases. 15 frontal faces of size 256×192 from the AR face database (men with and without glasses). Toy cars and toy animals of size 341×256 from the ETH 80 database (13 cars and 5 horses). We use segmentation maps to focus on the object features. Examples are shown in Figures 4 and 5.

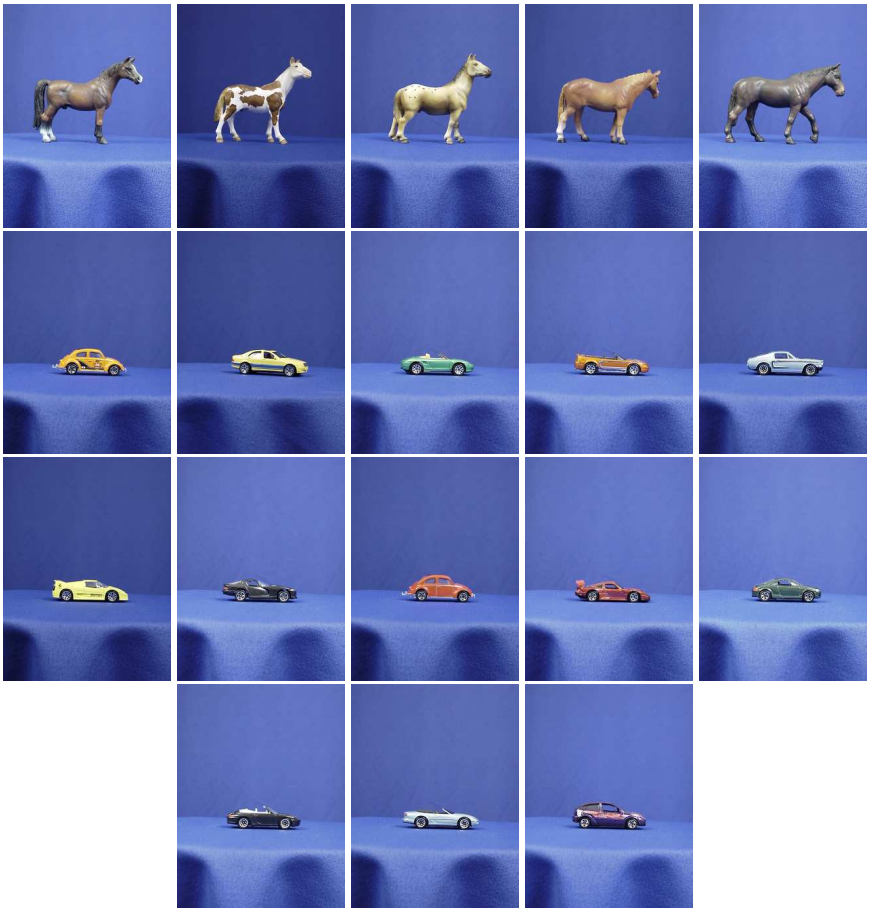

Fig. 4. Example of the ETH 80 database.

5.2 Experiments

This section shows the results for the databases. We only display classton channels that have high compactness and connectedness. The face example in Figure 5 demonstrates the robustness to scale changes and occlusions caused by facial hair or glasses. The example on the ETH 80 database shows that the classton channels are stable for visually similar views. In all our experiments, the classton channels of visually similar images produce the same spatial relations. These relations can be learned and the resulting model can then be used for classification. Such a model inherits the generalisation ability from the classtons and obtains discriminative power from the relative spatial relations.

The feature space used in the experiments is not normalised for orientation. The computed classtons are therefore orientation dependent. Orientation is an important feature for discrimination and in our examples orientation is needed to discriminate horizontal features from vertical features. If rotation invariance is required, than the invariance should be introduced by choosing a rotation invariant feature space as in [13].

Robustness to object identity We compute classtons from 15 frontal face images. Out of the 37 k-means clusters we choose the 4 classtons that mark the most significant features over several individuals (see Figure 5, corresponding clusters are marked with same grey level). Note, that nose, forehead, cheeks, eyes, chin, upper lips are marked by the same classton invariant of scale changes. The overall structure of the face is recovered even in the case of occlusions by facial hair, or glasses.

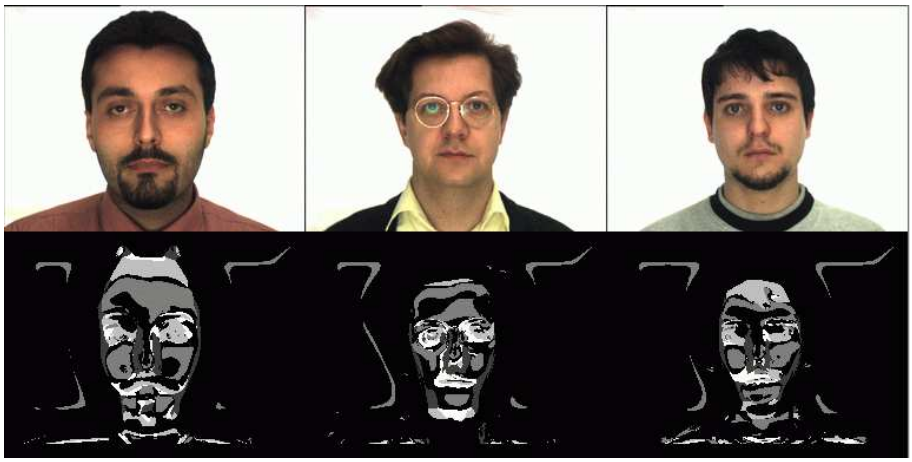

Fig. 5. Classton channel for frontal faces from the AR face database. Significant facial features are marked by the same classton channels, independent of identity, facial hair, glasses and scale change.

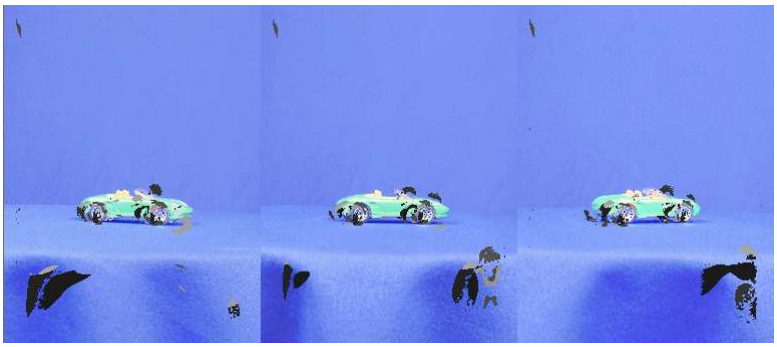

Fig. 6. Classton channels of the most compact and the most connected clusters computed from 26 side views of 13 toy cars (displayed are 6 channels coded as different grey levels). The classton channels are superposed on the original images. We observe the robustness to viewpoint changes. The wheels are automatically identified as significant features.

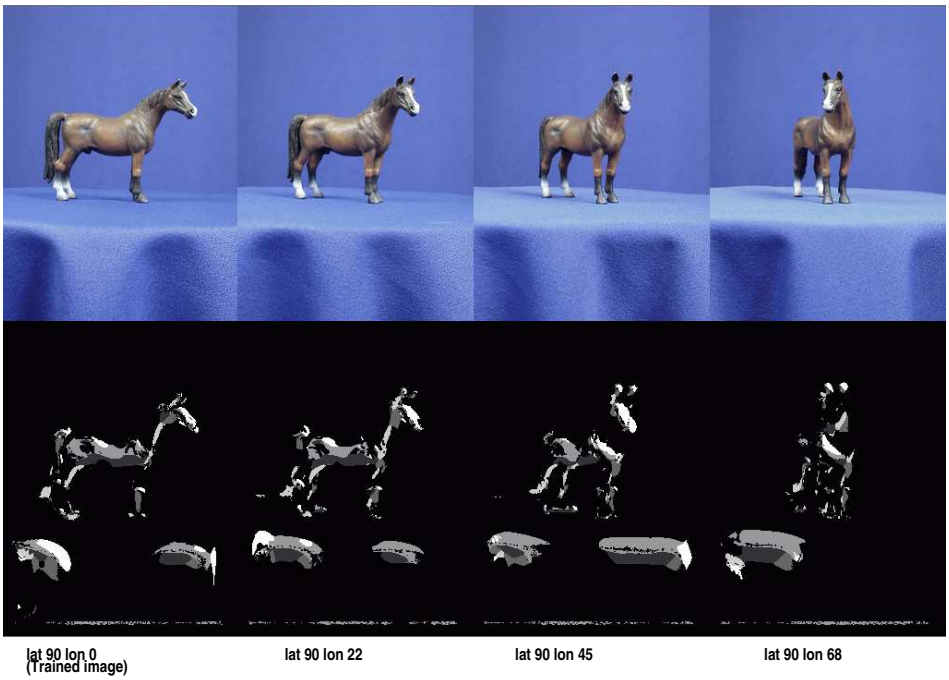

Fig. 7. Classton channels learned from full side view of 5 different horses. The channels are stable for visually similar views (here 45 degrees). Classification is possible. The channels are unstable for views that are not similar (right).

Robustness to viewpoint changes This section demonstrates the robustness to viewpoint changes on two examples. Figure 6 shows the 5 most compact and most connected classtons superposed on the image. We observe that the classtons determine the wheels as significant features among the set of 13 toy cars shown in Figure 4. Due to the very different car shapes, the wheels are the only common feature. Figure 7 shows an example of the robustness to viewpoint changes. Significant features are stable for visually similar views (the three leftmost figures). For views that are visually not similar (example Figure 7 right), the appearance of the channels changes considerably and classification would be difficult.

6 Conclusion

We propose a method to detect significant parts of the learned object robust to object identity, viewpoint, lighting conditions, pose, and scale of observation. Local appearance features are described by an appropriate feature space. Generic features are computed by unsupervised learning using clustering. The resulting classtons automatically identify significant class characteristics from a small set of examples. These significant characteristics are then reliably detected by means of the classton channels using a probabilistic cluster representation.

Local image features are often affected by noise. As a consequence, noise is present in the clusters computed from the features. We propose DBScan and k-means with pruning to reduce the sensitivity to noisy features. In order to select those classtons that display the best generalisation ability, we consider the density, the compactness of the clusters in feature space and the connectivity of the features in image space.

The reprojection of the clusters demonstrates generalisation ability with respect to previously unseen objects and robustness to viewpoint changes. The fusion of several such classton channels provides a powerful means for robust classification by preserving the internal scale relations of the class features.

Without the robust detection of significant class features any classification algorithm is going to fail. This is the motivation for this article and the presented clustering technique is a means to robustly detect and identify the significant class features which are essential for the composition of a model for classification.

The exact structure of the classification model is another complex problem and merits an article on its own. The classton approach that preserves internal scale relations of the features opens several interesting possibilities for the construction of classification models, among these scale invariant classification by shifting the class model in scale.

References

1. M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *International Conference on Knowledge Discovery and Data Mining (KDD-96)*, 1996.
2. M. Ester, H.-P. Krieger, J. Sander, M. Wimmer, and X. Xu. Incremental clustering for mining in a data warehouse environment. In *24th VLDB Conference*, New York, USA, 1998.
3. D.A. Forsyth and J. Ponce. *Computer Vision a Modern Approach*. Prentice Hall, 2003.

4. D. Hall, V. Colin de Verdière, and J.L. Crowley. Object recognition using coloured receptive fields. In *ECCV00*, Dublin, Ireland, June 2000.
5. J.J. Koenderink and A.J. van Doorn. Generic neighborhood operators. *PAMI*, 14(6):597–605, June 1992.
6. T. Leung and J. Malik. Recognizing surfaces using three-dimensional textons. In *ICCV*, Corfu, Greece, September 1999.
7. T. Lindeberg. Feature detection with automatic scale selection. *IJCV*, 30(2):79–116, 1998.
8. D.G. Lowe. Object recognition from local scale-invariant features. In *ICCV99*, pages 1150–1157, 1999.
9. J. Malik, S. Belongie, T. Leung, and J. Shi. Contour and texture analysis for image segmentation. *IJCV*, 43(1):7–27, June 2001.
10. A.M. Martinez and R. Benavente. The ar face database. Technical Report 24, CVC, June 1998.
11. R.P.N. Rao and D.H. Ballard. An active vision architecture based on iconic representations. *Artifi cial Intelligence*, 78(1–2):461–505, 1995.
12. B. Schiele and J.L. Crowley. Recognition without correspondence using multidimensional receptive field histograms. *IJCV*, 36(1):31–50, January 2000.
13. C. Schmid. Constructing models for content-based image retrieval. In *CVPR*, Kauai, USA, December 2001.
14. C. Schmid and R. Mohr. Local greyvalue invariants for image retrieval. *PAMI*, 1997.
15. S.-C. Zhu, C. Guo, Y. Wu, and Y. Wang. What are textons? In *ECCV02*, pages IV 793–807, 2002.