Semantic Grouping of Visual Features

Alexandra Teynor and Hans Burkhardt
Department of Computer Science, Albert-Ludwigs-Universität Freiburg, Germany
{*teynor, Hans.Burkhardt*}*@informatik.uni-freiburg.de*

Abstract

Many current object class models build on visual parts that constitute an object. However, visually different entities may actually refer to the same object part. This may be harmful for part based object class models. We present a method how visually distinct parts with the same semantic role can be associated by creating groupings based on the similarity of their occurrence distributions. Experimental results verify that more compact class representations can be built based on these groupings, which lead to improved classification performance and/or reduced classification time.

1. Introduction

A common technique for the recognition of object classes is the use of part dictionaries or "visual codebooks". These codebooks contain a variety of possible image structures. Whenever visual codebooks are created, e.g., by clustering appearance features from local image patches, we only have a *visual*, not a *semantic* grouping of object parts. The variety in the visual appearance of semantically equal object parts are due to several reasons. First, there are natural intra class variabilities. Then, we also have to deal with different poses, e.g., a mouth might be open, shut, or smiling showing the teeth. But also other reasons exist: current feature extraction methods often rely on interest point detections which are not always on the same locations on different object instances. This might result in shifted local windows for the same object part. So an eye might not always occur at the center of a local window, but also slightly shifted to the left or right. The features extracted from such shifted windows can be quite different. Invariance towards such shifts might be incorporated into the local features, but some very successful features like the SIFT features by Lowe [5] deliberately do not only consider the frequency of certain structures, but also their location. These types of features are affected by shifts in the detected structure.

Depending on the classification strategy, a separate treatment of semantically similar parts might be harmful. Especially when using "bag-of-feature" type approaches, parts with the same role are assigned to different dictionary entries. Distance calculation between part histograms is typically performed in a bin-by-bin fashion, so performance can be degraded by not relating semantically similar parts.

In this work, we present a novel way on how to perform a semantic grouping of object parts. Parts with a different visual appearance but with the same semantic role are associated by the similarity of their occurrence distributions given the object class.

2. Related work

Previous work concerning the semantic grouping of visual structures has been performed by Leibe [4] or Epshtein and Ullman [1]. Leibe combines visual parts by co-location and co-activation clustering. His approach is similar to ours as he also tries to associate parts that occur at the same location in an image, but he uses a weighted variation of the Hausdorff distance to combine visual parts. He does not apply his procedure to part frequency based object class models, as he advocates a Hough transform like voting method. Epshtein and Ullman use the context of parts in a probabilistic framework. They identify the geometric relation of parts co-occurring with a basic "root fragment", and search for similar constellations in test images. Our approach does not need a root fragment, but creates a number of groupings based on the desired similarity of the occurrence distributions.

3. Method

The basic idea is that object parts with the same semantic meaning occur at the same location(s) on an object. For example, the mouth is always located in

978-1-4244-2175-6/08/$25.00 ©2008 IEEE

$\mathbf{c}_i$ ω $p(x, y, s|\mathbf{c}_i, \omega)$

Figure 1. Creation of a part occurrence distribution.

the lower middle of a front view of a face, no matter whether it is surrounded by a mustache or laugh lines. So in order to learn semantic groupings, we need aligned training data, i.e. object parts with the same role should occur at the same location in an image.

We start from an initial codebook containing a variety of visual structures. The codebook can be acquired e.g. by clustering local appearance features from a training set. A general codebook $\mathbf{C}$ consists of a set of N vectors distributed in some feature space $\mathcal{X}$, which represent the individual clusters

$$\mathbf{C} = \{\mathbf{c}_i | \mathbf{c}_i \in \mathcal{X}, i = 1, \ldots, N\}. \quad (1)$$

Such a codebook now constitutes the possible visual appearances of object parts. In order to determine the semantic distance of object parts, we rely on the distribution of the occurrence of a certain structure given the object class ω. That is, for each of the clusters $\mathbf{c}_i$, a part occurrence distribution $p_i = p(x, y, s|\mathbf{c}_i, \omega)$ is built. x and y refer to the position of the occurrence of the structure, s to the scale. These density functions can be estimated for each cluster by matching the local features extracted from the aligned training data to the cluster centers. For each matching feature, the position and scale where it was extracted from is recorded in a histogram. The process is visualized in figure 1. On the left in this figure, enlarged sample members belonging to a cluster are displayed. In the middle, a sample image from the aligned training data is shown and on the right the resulting occurrence distribution at a particular scale.

We define the semantic distance of two clusters to be the distance of their distribution maps

$$d_{\text{sem}}(\mathbf{c}_i, \mathbf{c}_j) = d(p_i, p_j). \quad (2)$$

For d, several functions can be used. In this work, we use normalized cross correlation, but other measures could be chosen as well. Since correlation is a similarity measure, not a distance measure (the value becomes bigger when the vectors are more similar), we subtract the value from the maximal possible value for normalized correlation, which is 1. We use a non parametric representation for the location distributions in the form of a three dimensional histogram (x-location, y-location and scale). Let $\mathcal{H}$ and $\mathcal{R}$ be normalized histograms with N bins, so that $\sum_{i=1}^{N} h_i = \sum_{i=1}^{N} r_i = 1$. h_i and r_i represent the individual bin values. Then, the normalized cross correlation distance is

$$d_{\text{corr}}(\mathcal{H}, \mathcal{R}) = 1 - \frac{\sum_{i=1}^{N} h_i r_i}{\sqrt{\sum_{i=1}^{N} h_i^2} \sqrt{\sum_{i=1}^{N} r_i^2}}. \quad (3)$$

We can now determine whether two clusters $\mathbf{c}_i$ and $\mathbf{c}_j$ are semantically related by an indicating function $f_{\omega,\vartheta}(\mathbf{c}_i, \mathbf{c}_j)$, given the class ω and a threshold ϑ in the following way:

$$f_{\omega,\vartheta}(\mathbf{c}_i, \mathbf{c}_j) = \begin{cases} 1 & d_{\text{sem}}(\mathbf{c}_i, \mathbf{c}_j) < \vartheta \\ 0 & \text{otherwise} \end{cases}. \quad (4)$$

It is important to note that we do not deal with equivalence classes here, since the transitivity condition is not met. It can easily be seen that if $f_{\omega,\vartheta}(\mathbf{c}_i, \mathbf{c}_k) = 1$ and $f_{\omega,\vartheta}(\mathbf{c}_k, \mathbf{c}_j) = 1$, it does not necessarily follow that $f_{\omega,\vartheta}(\mathbf{c}_i, \mathbf{c}_j) = 1$, since the similarity between their occurrence distributions does not need to be below the given threshold. This problem is also present in natural languages: pairs of words that can be used synonymously also do not have to follow transitivity conditions, e.g. the words sick↔bad as well as bad↔evil can be used interchangeably, however the words sick↔evil are not directly related any more.

In the following, we want to investigate the effect of grouping semantically similar parts together in a bag-of-feature classification approach. Initially, we have a part histogram $\mathcal{H}$ of local structures in dimension N, i.e. the number of clusters in the codebook. The bin counts h_i are determined by membership functions $w_i(\mathbf{x})$, $\sum_{i=1}^{N} w_i(\mathbf{x}) = 1$ for bin i

$$h_i = \sum_{l=1}^{L} w_i(\mathbf{x}_l), \quad (5)$$

where $\mathbf{x}_l \in \mathcal{X}, l = 1, \ldots, L$ are local feature vectors extracted from an image. The membership function might be binary, only assigning the feature vector to the nearest codebook entry, or more complex, distributing the count between several entries. The contributions of the individual local feature vectors are accumulated in the histogram and the histogram renormalized to add up to one. It is important to note that w_i deals with the *visual* similarity here.

We then have to decide which visually distinct, but semantically similar parts should be treated together. We do not directly use the semantic indicating function $f_{\omega,\vartheta}(\mathbf{c}_i, \mathbf{c}_j)$ due to the transitivity problem mentioned

above. Instead, we cluster the elements of our visual part dictionary, using the semantic similarity measure as defined in equation (2) and (3). In our experiments, we use agglomerative clustering [2] with an average link paradigm, since it produces compact clusters and only relies on pairwise similarities of feature vectors. The final groupings can be determined by cutting the hierarchical tree at an appropriate value. For each visual part i, we then obtain an index $s_i \in \{1, \ldots, M\}$ describing the semantic cluster membership.

We now want to combine the part frequencies of an original histogram $\mathcal{H}$ that belong to the same semantic cluster in order to obtain a more general part histogram $\mathcal{M}$ with reduced dimensionality M. The new histogram entries are given by

$$m_k = \sum_{i=1}^{N} h_i \delta(s_i, k),\ k \in 1, \ldots, M \qquad (6)$$

where δ is the Kronecker delta function. In effect, all entries of the original histogram that are in the same semantic cluster are added together.

4. Experiments

In order to show the benefits of semantic grouping of visual distinct structures, we performed various experiments for different object categories, in particular, natural as well as mechanical objects. As we need aligned training data for building the location maps, we chose the categories "Faces_easy" and "Motorbikes" from the Caltech101 dataset[1] for our experiments, as they fulfill this condition very well. In the faces dataset, there are 435 images of 31 different people, the motorbikes dataset consists of 789 images from different motorbikes. We used half of the images for training, the other half for testing. As negative examples for training and testing, the respective same number of images was drawn randomly from the remaining object categories.

Only gray scale information was used for feature extraction. We calculated GLOH features around Hesse-Laplace interest points, as they have proven to be successful in a comparative study [6] as well as in our own experience. We used the original detector and descriptor code from the authors.

The codebook describing the visual structures under consideration was obtained by clustering 100000 local features randomly selected from the training images. The MBSAS clustering scheme [7] was used for this purpose, as it allows clustering a large number of features in reasonable time. Then histograms based on the basic codebook were calculated using a sigmoid matching function.

[1] from http://www.robots.ox.ac.uk/~vgg/data3.html

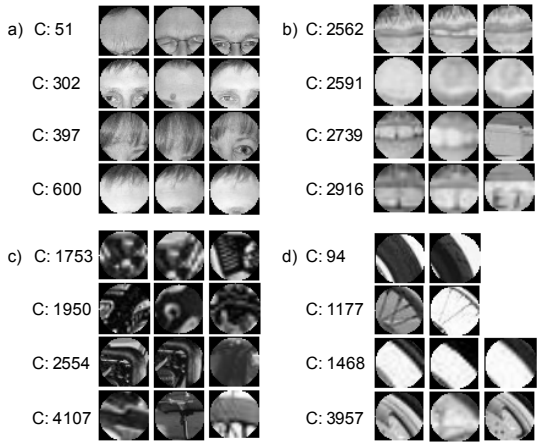

Figure 2. Examples of semantic groupings for the Caltech classes "faces" and "motorbikes". Each row shows examples of a visual cluster.

In order to give an idea about the type of visual structures combined by our process, we show sample semantic groupings in figure 2. We can visually verify that combining these structures is sensible. For each semantic grouping, 4 visual clusters are shown. A visual cluster is represented by a row: the codebook number followed by at maximum 3 sample patches from the visual cluster. So the horizontal direction shows visual groupings, the vertical direction semantic groupings. We can see for the class "faces": a) foreheads with different hair styles, b) mouth parts; for the class "motorbikes": c) engine parts and d) wheel parts.

As a sample task, we deal with a two class problem where the presence of a member of a specific object class in an image should be determined. In order to show the effect of semantic clustering on different types of classifiers, we use a simple k-nearest neighbor classifier with histogram intersection as a distance measure, with k = 3 in our case, as well as a SVM with an histogram intersection kernel. We test the behavior for the classifiers according to different cut values in the semantic clustering step.

In figure 3, we list the results for the Motorbikes class. The results for the faces class are very similar and omitted due to space reasons. In the top of figure 3, the dimensionality of the combined clusters is shown for different cut values. It depends on the number of initial clusters, the number of clusters that were combined and the number of semantic clusters currently established. In the bottom, we can see the results of the

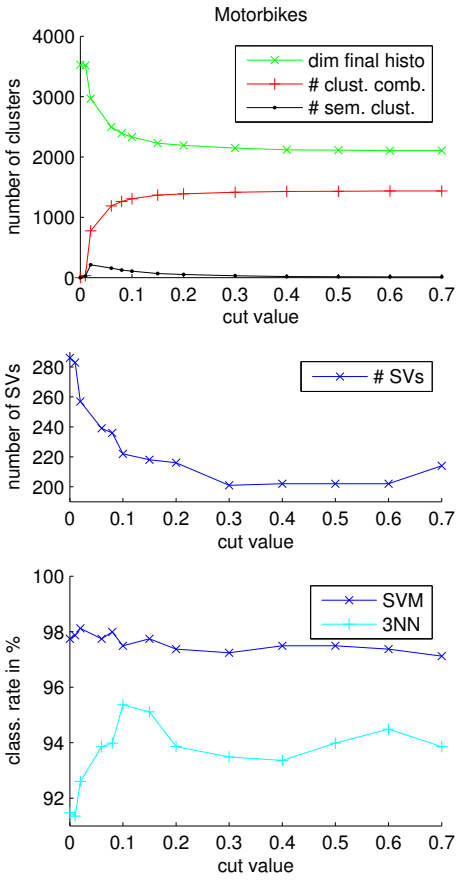

Figure 3. Results for the experiments. For a detailed explanation, see the text.

classification task for the different classifiers. We can see that of course the SVM based results are superior to the NN-classifiers due to the generalization capabilities of the large margin classifier. Nevertheless we show the results for the NN classifier, since an improvement in the quality of the features can be seen more directly there. And indeed, for the NN classifier, the classification performance increases, e.g. from 91.5% for the original histogram to 95.4% for the semantically combined histogram for the motorbikes class.

The SVM cannot profit that much from the classification accuracy point of view. It even drops slightly (from 97.7% to 97.5% for the same cut values). However the number of support vectors decreases which means that we have a more "simple" decision boundary in the mapped feature space. We could save up to 14.3% of the support vectors (from 209 to 179) for the faces class with no loss in performance. Together with the reduced dimensionality of the feature vectors, this means less classification time.

For too low cut values, where many relevant parts were mapped to few histogram entries, classification performance drops for the NN classifier and the number of support vectors increases again. The specific cut values for grouping features must be determined experimentally and can be estimated using a validation set. Generally rather low cut values are already sufficient to improve performance.

5. Conclusion

In this work, we presented a method on how to establish a semantic grouping for object parts depending on the similarity of their occurrence distributions. In this way, parts that are visually distinct, but semantically similar can be associated and processed together.

We have shown that for object class representations based on part histograms (bag-of-feature type approaches) it is beneficial to associate semantically similar parts by combining the respective visual clusters. Simple classifiers like a nearest neighbor classifier can directly profit from that. Also more complex classifiers like SVMs, that were able to deal with the diversity of semantically similar object parts in the first place, can also benefit from the process by a reduced set of support vectors and a smaller feature dimension which lead to a decrease in classification time.

The semantic mapping step can be easily incorporated into other powerful classification approaches where bag-of-feature representations are involved, like e.g. the spatial pyramid matching by Lazebnik et al. [3]. The approach is easy to implement yet very effective.

References

[1] B. Epshtein and S. Ullman. Identifying semantically equivalent object fragments. In *Proc. CVPR*, volume 1, pages 2–9, 2005.

[2] A. Jain and R. Dubes. *Algorithms for clustering data*. Prentice-Hall, Inc. Upper Saddle River, NJ, USA, 1988.

[3] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. CVPR*, volume 2, pages 2169–2178, 2006.

[4] B. Leibe. *Interleaved Object Categorization and Segmentation*. PhD thesis, ETH Zurich, 2004. PhD Thesis No. 15752.

[5] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60:91–110, 2004.

[6] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE TPAMI*, 27(10):1615–1630, 2005.

[7] A. Teynor and H. Burkhardt. Fast codebook generation by sequential data analysis for object classification. In *Proc. ISVC*, Lake Tahoe, USA, 2007.