The 22nd International Conference on Machine Learning

7-11 August 2005 in Bonn, Germany

W 9

Proceedings of the Workshop on

Machine Learning Techniques for Processing Multimedia Content

Matthieu Cord

Pádraig Cunningham

Rozenn Dahyot

Tamas Sziranyi

ICML 2005 BONN - GERMANY

Preface

Machine Learning (ML) techniques are used in situations where data is available in electronic format and ML algorithms can 'add value' by analysing this data. This is the situation with the processing of multimedia content. The 'added value' from ML can take a number of forms:

- by providing insight into the domain from which the data is drawn,
- by improving the performance of another process that is manipulating the data,
- by organising the data in some way or
- by helping to interpret multimedia content to make it more understandable.

This potential for ML to add value in processing of multimedia content has made this one of the most popular application areas for ML research. Multimedia content has some characteristics that place specific demands on ML. The data is typically of very high dimension and dimension reduction is often required. The normal distinction between supervised and unsupervised techniques doesn't always apply; it is often the case that only some of the data is labeled or the user may assist in labeling the data during processing. Typically the ML process is preceded by a feature extraction stage and the success of the ML stage will often depend on the feature extraction.

This workshop on Machine Learning Techniques for Processing Multimedia Content has been organized because of these special issues that arise with multimedia data. We have papers describing applications in image processing, video analysis and music classification. The research described in these papers has drawn on a wide range of ML techniques. It is hoped that this workshop will help identify important research directions for Machine Learning that will help in the processing of multimedia content.

We would like to express our thanks to the Programme Committee for their help in selecting the papers for presentation at this workshop. Finally, we thank Hendrik Blokeel for his overall organization of the 2005 ICML workshop series.

June 2005

Matthieu Cord
Pádraig Cunningham
Rozenn Dahyot
Tamás Szirányi

Workshop Organisation

Co- chairs

Matthieu Cord
Pádraig Cunningham
Rozenn Dahyot
Tamás Szirányi

Programme Committee

Horst Bischof, Graz University of Technology Austria
Matthieu Cord, ENSEA France
Pádraig Cunningham, Trinity College Dublin Ireland
Rozenn Dahyot, Trinity College Dublin Ireland
Christophe Garcia, France Telecom R&D France
Christophe Laurent, France Telecom R&D France
Nathalie Laurent, France Telecom R&D France
Shaul Markovitch, Technion Israel
Eric Moulines, ENST France
Eric Pauwels, CWI The Netherlands
Nicu Sebe, University of Amsterdam The Netherlands
Ovidio Salvetti, CNR Italy
Fred Stentiford, UCL United Kingdom
Tamás Szirányi, SZTAKI Hungary
Dietrich Wettschereck, Recommind Germany

Contents

Motion Analysis and Synthesis of Time-dependent Data

Hongyu Li HONGYULI@FUDAN.EDU.CN
Shanghai Key Laboratory of Intelligent Information Processing (IIPL),
Department of Computer Science and Engineering, Fudan University, Shanghai, 200433, China

Wenbin Chen WBCHEN@FUDAN.EDU.CN
Department of Mathematics, Fudan University, Shanghai, 200433, China

I-Fan Shen YFSHEN@FUDAN.EDU.CN
Su Yang SUYANG@FUDAN.EDU.CN
Shanghai Key Laboratory of Intelligent Information Processing (IIPL),
Department of Computer Science and Engineering, Fudan University, Shanghai, 200433, China

Abstract

In this paper temporal local tangent space alignment is proposed to deal with time-dependent data, such as video and motion capture data. It is an extension of local tangent space alignment, for short, LTSA, from spacial to temporal learning. LTSA is a nonlinear dimension reduction method based on Euclidean distance. Temporal LTSA, however, is dependent on the continuity of time of input data. Another algorithmic improvement is made upon LTSA for mapping new data between the low- and high-dimensional spaces, which makes LTSA suitable in a changing, dynamic environment. When temporal LTSA is applied to time-dependent data, motion of objects underlying in such data can be carefully analyzed in a low-dimensional space. Motion decomposition and synthesis can be further made for real applications.

1. Introduction

Time-dependent data are those data containing information about time continuity such as video and motion capture data. The variation of such data is related with time. The raw time-dependent data taken with cameras or other capturing devices are in general of very high dimensionality. In nature, however, only a few degrees of freedom play an important role in the process of real human or animal motion. For example, one of most prominent features of human walking is the periodicity. For the convenience of studying the periodicity, human motion can be described using one degree of freedom which cyclically varies with time. Thus dimension reduction is necessary.

Dimension reduction is a preprocessing step for analysis of high-dimensional time-dependent data and acts as an important role to synthesize smoother and more continuous movement. Traditional methods to perform dimension reduction are mainly linear, including principal component analysis and multidimensional scaling (Duda et al., 2001).

Recently, a conceptually simple yet powerful method for nonlinear dimension reduction has been proposed in (Zhang & Zha, 2004): local tangent space alignment (LTSA). Its basic idea is that the global structure of a nonlinear manifold can be obtained from the interaction of overlapping local tangent spaces. LTSA is superior to another popular nonlinear mapping method, locally linear embedding (LLE) (Roweis & Saul, 2000) since the LTSA method is able to discover more useful degrees of freedom than the LLE method (Li et al., 2005).

Although the authors demonstrate their algorithm on a number of artificial and realistic data sets, there have as yet been few reports of application of LTSA. LTSA does not derive an explicit mapping function between the high- and low-dimensional spaces, therefore when new data arrive, we have to put all data together and compute again, i.e., LTSA is stationary with respect to data and lacks generalization to new data. In this paper, this problem will be addressed. We propose a simple technique to map new data in the high- or low-dimensional space to another space, which makes LTSA suitable in a changing, dynamic environment. Besides, temporal LTSA (TLTSA) is specially proposed for dealing with time-dependent data where the

Appearing in *Proceedings of the workshop Machine Learning Techniques for Processing Multimedia Content*, Bonn, Germany, 2005.

measure of neighborhood selection is not Euclidean distance, but time interval.

The remainder of the paper is divided into the following parts. In section 2 LTSA is briefly introduced and extended to adapt itself to a dynamic environment. Section 3 presents the temporal variation of LTSA. Some experimental results are presented in section 4. Finally, section 5 ends with some conclusions.

2. Local Tangent Space Alignment

Time-dependant data can be essentially considered as a sequence of vectors. For example, a fraction of video is composed of many frames of images considered as points with coordinate vectors in a high-dimensional image space.

Let us consider a set of input points with coordinate vectors $X = \{x_i\}_{i=1}^n$ in R^m. Our aim is to obtain a set of output vectors $Y = \{y_i\}_{i=1}^n$ in a *d*-dimensional space where $d < m$. In this paper, we use local tangent space alignment (LTSA) to achieve this goal. It assumes that all data lie on or close to a nonlinear manifold and the global geometrical structure of this manifold can be learned by analyzing its overlapping local geometrical structure. It treats local tangent space of each point as such geometry and aligns those tangent spaces between the high- and low-dimensional spaces. The corresponding low-dimensional coordinates Y are discovered in the process of alignment.

2.1. Summary of LTSA

In this paper we will not discuss the derivation and proof for LTSA (For details, please refer to (Zhang & Zha, 2004)).

Next we briefly describe how to extract low-dimensional coordinates Y from a set of high-dimensional data X with LTSA.

1. Find k nearest neighbors $X_i = \{x_i^j\}, j = 1, \ldots, k$ for each point $x_i, i = 1, \ldots, n$.

2. Extract the local geometrical information by calculating the d largest eigenvectors $g_1, \ldots, g_d$ of the correlation matrix $(X_i - \bar{x}_i e^T)^T (X_i - \bar{x}_i e^T)$. e is a column vector whose entries are all ones. $\bar{x}_i$ represents the average of the neighborhood of x_i, $\bar{x}_i = \frac{1}{k} \sum_j x_i^j$. Set $G_i = [e/\sqrt{k}, g_1, \ldots, g_d]$.

3. Construct the alignment matrix B by locally summing as follows:

$$B(I_i, I_i) \leftarrow B(I_i, I_i) + I - G_i G_i^T, i = 1, \ldots, n$$

with initial $B = 0$. I is a $k \times k$ identity matrix, I_i denotes the set of indices for the k nearest neighbors of x_i.

4. Compute the $d+1$ smallest eigenvectors of B and pick up the eigenvector matrix $[u_2, \ldots, u_{d+1}]$ corresponding to the 2nd to d+1st smallest eigenvalues. Set the global coordinates

$$Y = [y_1, \ldots, y_n] = [u_2, \ldots, u_{d+1}]^T.$$

Note neighborhood selection (the first step of LTSA) to estimate the local tangent space is very crucial to the success of this algorithm. In the simplest formulation of the algorithm, one identifies a fixed number of nearest neighbors, k, per data point, as measured by Euclidean distance. Fig.1 shows the selection of neighbors in a 3-D space. All data points discretely distribute on the surface of a ball. Yellow squares surrounded with a black loop are 15 nearest neighbors of x_i. They together with x_i compose a local neighborhood of x_i. Two of tangent vectors at x_i, T_1 and T_2, approximately span the tangent space of x_i.

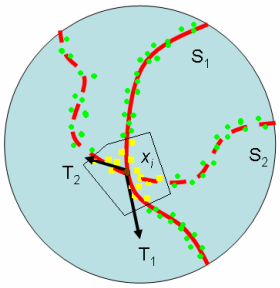


Figure 1. The selection of nearest neighbors in the 3-D Euclidean space. Yellow squares surrounded with a black loop are 15 nearest neighbors of x_i.

Other criteria, however, can also be used to choose neighbors. For example, one can identify neighbors by choosing all points within a ball of fixed radius. One can also use locally derived distance metrics based on a priori knowledge such as class membership or time order, which deviate significantly from a globally Euclidean norm. In general, neighborhood selection in LTSA presents an opportunity to incorporate a priori knowledge.

2.2. Dynamic LTSA

The original LTSA is stationary with respect to the data, that is, it requires a whole set of points as an input in order to map them into the embedding space. When new data points arrive, the only way to map them is to pool both old and new points and return LTSA again for this pool. Therefore, the original LTSA lacks generalization to new data, it is not suitable in a changing, dynamic environment.

Our attempt is to adapt LTSA to a changing situation where the data come incrementally point by point, and avoid an expensive eigenvector calculation for each new query,

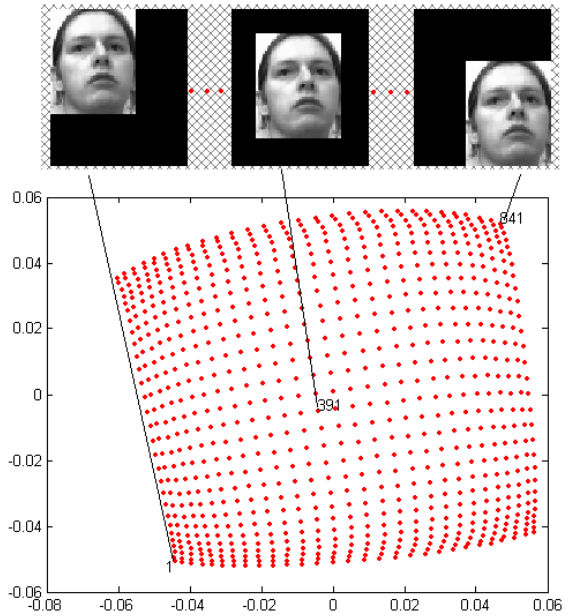


Figure 2. Successful recovery of a manifold of known structure using LTSA.

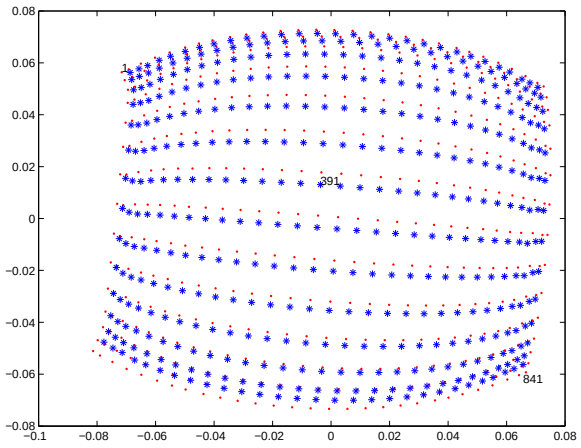


Figure 3. Mapping new arriving test images to the 2-D feature space learned from training images.

which is inspired from (Saul & Roweis, 2002; Kouroteva et al., 2002). Let a set of points $X = \{x_i\}$, $i = 1, \ldots, n$, as an original input to LTSA. After dimension reduction with LTSA, the projection of X to the embedding space can be discovered, $Y = \{y_i\}$, $i = 1, \ldots, n$. In particular, to compute the output y_{n+1} for a new arriving point x_{n+1}, we can do the following. First look for the point x_j closest to x_{n+1} among X. Let y_j be the projection of x_j to the embeddding space. The derivation of LTSA reveals that the following equation is approximately true:

$$x_j - \bar{x}_j \approx J_f(y_j - \bar{y}_j) = T(y_j - \bar{y}_j),$$

where T is a transformation matrix of size $m \times d$, $\bar{x}_j$ and $\bar{y}_j$ are respectively the mean of k nearest neighbors of x_j and y_j. The matrix T can be straightforwardly determined as

$$T = (x_j - \bar{x}_j)(y_j - \bar{y}_j)^+. \quad (1)$$

where $(\cdot)^+$ represents the Moore-Penrose generalized inverse of a matrix. Assume x_j and x_{n+1} lie close enough to each other (Note that input data X must be dense enough to sufficiently cover the whole surface of the embedded manifold, or else our method of generalization can not perform well), so the transformation matrix T of x_j is applicable to x_{n+1}. y_{n+1} can be obtained

$$y_{n+1} = \bar{y}_j + T^+(x_{n+1} - \bar{x}_j). \quad (2)$$

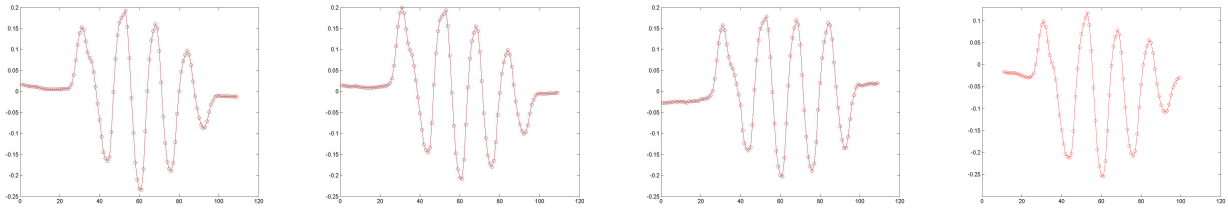
A mapping from the embedding space to the input space can also be derived in the same manner.

Let us consider 841 grayscale images of a single face translated across a two-dimensional background shown in the top panel of Fig.2. Such images lie on an intrinsically two-dimensional nonlinear manifold in bottom panel, but have an extrinsic dimensionality equal to the number of pixels in each image (m=2576).

We divide the 841 images into two sets, which are alternately selected from left to right along the horizontal direction. One set including 435 images is considered as original input data to learn a manifold, the other including 406 images as new arriving data to test dynamic LTSA. Similar to the one using LTSA, the result using dynamic LTSA shown in Fig.3 successfully maps the images with corner faces to the corners of its two dimensional embedding and does reflect the character of face movement. But dynamic LTSA only spends about 32.5 seconds on computation, which improves the efficiency by 40% in comparison with 53.9 seconds that LTSA needs. Note the advantage of our method is not obvious when the test data set is too small, but when such set becomes very large, the computation efficiency will be greatly improved if using our method of generalization .

3. Temporal LTSA

As stated in section 2.1, the criterion of neighborhood selection is supposed to the crucial feature of data sets. For time-dependent data, time order is more important than Euclidean distance, so in the process of neighborhood selection, we can employ time order to decide the nearest neighbors of each point. This improved method is called temporal LTSA (TLTSA). For example, in Fig.5, if we take $k = 5$, the nearest neighbors of the i-th data along the time axis are those points between $i - 5$-th and $i + 5$-th.



(a) Extrapolation (b) Adding-zero (c) Overlapping (d) Truncation

Figure 4. Four feasible methods for dealing with margin points. a: the extrapolation method, b: the adding-zero method, c: the overlapping method, d: the truncation method.

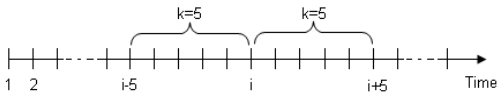


Figure 5. The selection of nearest neighbors for time-dependent data.

For some points at the beginning and end of the sequence, their neighborhood is out of the sequence, thus the nearest neighbors can be chosen in some special ways. If not adding new data to the sequence, we can directly select some points near by them which may lead to the same neighborhood for some margin points. Or, we increase new data in order at the beginning and end of the sequence. New data can be same or not, for instance, they can be all-zero vectors; or extrapolate in terms of margin points to obtain new points. Another method is to give up those margin points to consider. In nature, all these methods have the similar effects.

Fig.4 presents four feasible methods for dealing with margin points. This set of data taken with motion capture devices represent the walking motion of a human and will be in detail introduced in section 4.2. Since the walking motion is approximately periodical, such data should essentially contain a significant degree of freedom which varies approximately periodically with time. The original data are of 54 dimensions, now we map them to a 1-D space with temporal LTSA. Four different methods are used for neighborhood selection of margin points. The extrapolation method first generates new data according to the original data and then inserts them on both sides of the sequence in order. The adding-zero method adds all-zero vectors on the outside of the sequence. The overlapping method does not introduce new data, it only admits margin points of a sequence to use the common neighborhood. The truncation method directly gives up those margin points and only considered those points with applicable neighborhood.

We have to again stress that the goal of the temporal LTSA algorithm is to obtain the global geometrical feature of a set of data by the analysis of its local tangent space constructed in terms of time order. Experimental results show that the algorithm is not sensitive for local discontinuity of data, i.e., adding/deleting some data to estimate local tangent spaces will not actually damage the character of the global geometry. In our following experiments, we will mainly use the overlapping method.

Besides, we have found that if the number of nearest neighbors k is set too small, the mapping will not reflect any global properties of data; if it is too high, the mapping will lose its nonlinear character and behave like traditional PCA, as the entire data set is seen as local neighborhood. The algorithm is stable over a wide range of values but do break down as k becomes too small or large.

4. TLTSA for Motion Analysis and Synthesis

In general, time-dependent data only can be shown by some special softwares or player. We can not directly take a complete view of their global continuity and smoothness. For better analyzing motion contained in time-dependent data, the TLTSA algorithm can be used to map these data to a lower-dimensional space (exactly 1-, 2- or 3-dimensional). Next we will respectively discuss two classes of time-dependent data, video and motion capture data, and present some experimental results.

4.1. Video Data

Video is composed of a sequence of images. If each image is represented by a vector, a section of video can be considered as a set of vector data with time order. Such data are high-dimensional, which is not beneficial for our direct analysis. When moving objects exist in a video, the vector data necessarily contain related motion information. Since most realistic movements such as walking, running and jumping are of low degrees of freedom, the high-dimensional vector data to represent this section of video contain a lot of redundant or insignificant information. Eliminating them can make it easy to study the moving feature of such data. Here we apply the TLTSA algo-

rithm into 35 face images (Fig.6) composing a section of discontinuous video about face rotation. A 2-D embedding was discovered by TLTSA and shown in Fig.7 where blue stars correspond to input training face images. Linear interpolation is made between each pair of sequential points in the low-dimensional space and the results are represented by red points of Fig.7. By mapping those newly inserted points from the low- to high-dimensional spaces, one can obtain some new face images in the high-dimensional image space. Insert these new images back to the original image sequence, a section of new smooth video about face rotation can be generated.

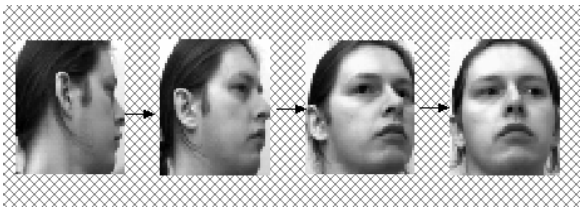


Figure 6. 4 of 35 face images taken from different view angles – from side to frontal.

4.2. Motion Capture Data

Motion capture data belong to another class of time-dependent data, which are widely applied in the field of character animation. Such data are not contaminated by the variation of background, therefore they can better embody the feature of motion than video data. Here we use a set of data with 54 dimensions representing walking of virtual human (Fig.8) to analyze basic properties of human walking. Since human walking is or close to periodical, there should exist a degree of freedom describing such periodicity. Map such 109 continuous data to a 1-D space with TLTSA and show their variational regularity with time in Fig.9. Each red point corresponds to a walking state of virtual human and some key states corresponding to blue stars have be shown around the chart. Fig.9 reveals that the whole process of human walking can be approximately divided into three stages: the beginning (from Time 1 to Time 23), the circular advancement (from Time 24 to Time 99) and the ending (from Time 100 to Time 109). The advancement stage is composed of four primitive cycles. Extraction of the primitive motion from this stage allows us to arbitrarily copy such primitive and synthesize similar motion sequences. That is, one can randomly assign the number of cycle and let virtual human smoothly walk in terms of the assigned number.

5. Conclusions

In this paper, we propose a simple technique to map new data in the high- or low-dimensional space to another space,

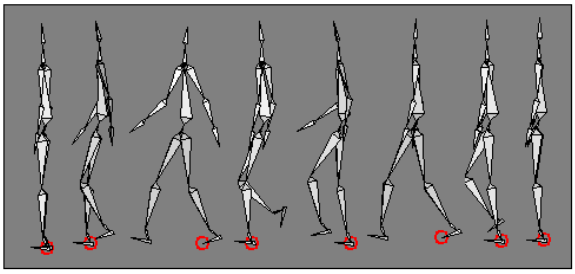


Figure 8. The walking process of virtual human which is used to capture motion data.

which makes LTSA suitable in a changing, dynamic environment. Besides, temporal LTSA is specially proposed for dealing with time-dependent data where the measure of neighborhood selection is not Euclidean distance, but time interval. Experiments show that the TLTSA algorithm can efficiently extract key degrees of freedom from time-dependent data, which is very beneficial for motion analysis and synthesis.

Acknowledgments

This work was supported by NSFC under contract 60473104 and STCSM under contract 045115013.

References

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification*. New York, NY: John Wiley & Sons. 2 edition.

Kouroteva, O., Okun, O., Soriano, M., Marcos, S., & Pietikainen, M. (2002). *Beyond locally linear embedding algorithm*Technical Report MVG-01-2002). Machine Vision Group, University of Oulu, Finland.

Li, H., Chen, W., & Shen, I.-F. (2005). Supervised learning for classification. *ICNC'05-FSKD'05, LNCS (to appear)*. Springer-Verlag.

Roweis, S., & Saul, L. (2000). Nonlinear dimension reduction by locally linear embedding. *Science, 290*, 2323–2326.

Saul, L., & Roweis, S. (2002). *Think globally, fit locally: unsupervised learning of nonolinear manifolds*Technical Report MS CIS-02-18). Univ. Pennsylvania.

Zhang, Z., & Zha, H. (2004). Principal manifolds and nonlinear dimension reduction via local tangent space alignment. *SIAM Journal of Scientific Computing, 26*, 313–338.

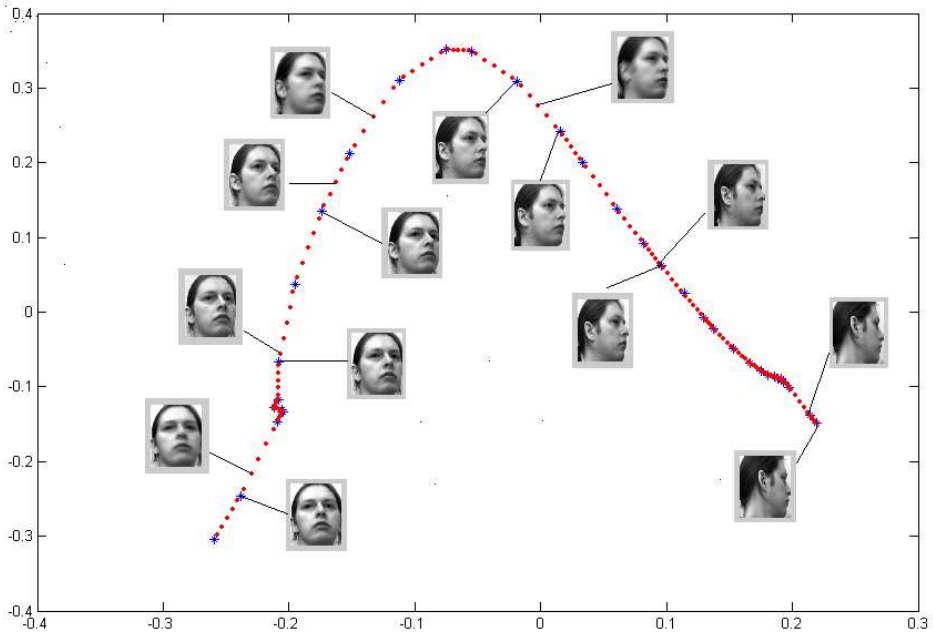


Figure 7. The mapping between the two- and high-dimensional spaces with TLTSA. Blue stars are generated from the input training face images. Red points are obtained by linear interpolation.

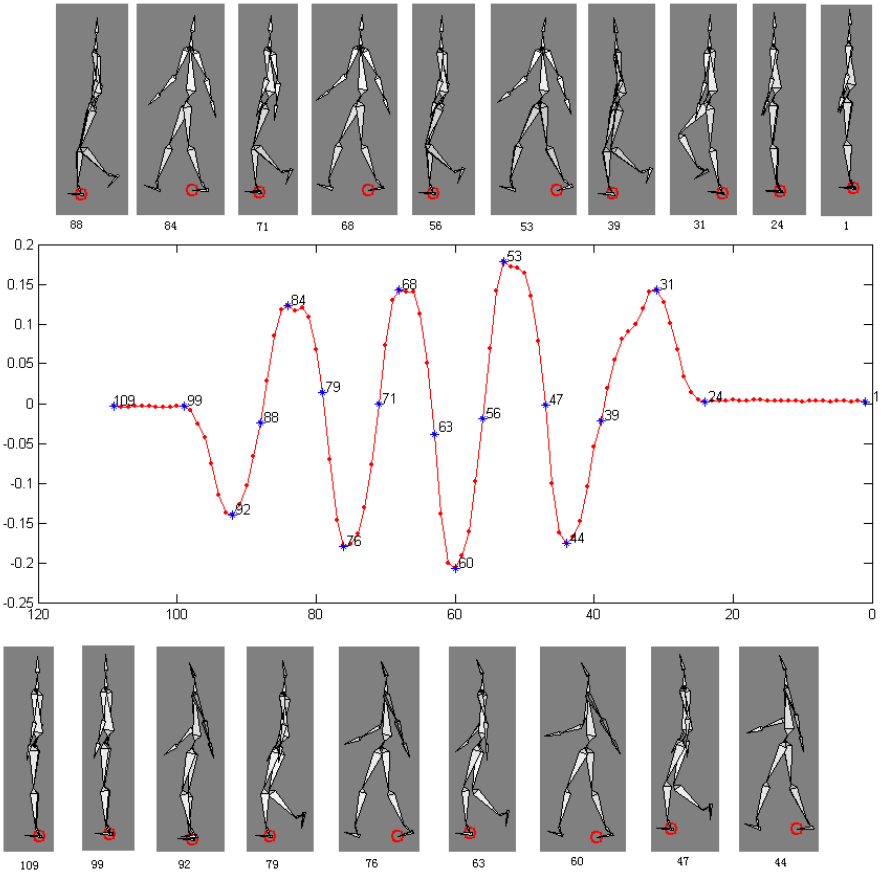


Figure 9. Periodicity of human walking. The horizontal axis represents the variation of time and gradually increases from right to left.

Decision Trees and Random Subwindows for Object Recognition

Raphaël Marée RAPHAEL.MAREE@ULG.AC.BE
Pierre Geurts PIERRE.GEURTS@ULG.AC.BE
Justus Piater JUSTUS.PIATER@ULG.AC.BE
Louis Wehenkel LOUIS.WEHENKEL@ULG.AC.BE
Department of EE & CS, Institut Montefiore, University of Liège, B-4000 Liège, Belgium

Abstract

In this paper, we compare five tree-based machine learning methods within our recent generic image-classification framework based on random extraction and classification of subwindows. We evaluate them on three publicly available object-recognition datasets (COIL-100, ETH-80, and ZuBuD). Our comparison shows that this general and conceptually simple framework yields good results when combined with ensembles of decision trees, especially when using Tree Boosting or Extra-Trees. The latter is particularly attractive in terms of computational efficiency.

1. Introduction

Object recognition is an important problem within image classification, which appears in many application domains. In the object recognition literature, local approaches generally perform better than global approaches. They are more robust to varying conditions because these variations can locally be modelled by simple transformations (Matas & Obdržálek, 2004). These methods are also more robust to partial occlusions and cluttered backgrounds. Indeed, the correct classification of all local features is not required to correctly classify one image. These methods are generally based on region detectors (Mikolajczyk et al., 2005) and local descriptors (Mikolajczyk & Schmid, 2005) combined with nearest-neighbor matching.

In this paper, we compare five tree-based machine learning methods within the generic image classification framework that we proposed in earlier work (Marée et al., 2005). It is based on random extraction of subwindows (square patches) and their classification by decision trees.

Appearing in *Proceedings of the workshop Machine Learning Techniques for Processing Multimedia Content*, Bonn, Germany, 2005.

2. Framework

In this section, we briefly describe the framework proposed by (Marée et al., 2005). During the training phase, subwindows are randomly extracted from training images (2.1), and a model is constructed by machine learning (2.2) based on transformed versions of these (Figure 1). Classification of a new test image (2.3) similarly entails extraction and description of subwindows, and the application of the learned model to these subwindows. Aggregation of subwindow predictions is then performed to classify the test image, as illustrated in Figure 2. In this paper, we evaluate various tree-based methods for learning a model.

2.1. Subwindows

The method extracts a large number of possibly overlapping, square subwindows of random sizes and at random positions from training images. Each subwindow size is randomly chosen between 1×1 pixels and the minimum horizontal or vertical size of the current training image. The position is then randomly chosen so that each subwindow is fully contained in the image. By randomly selecting a large number (N_w) of subwindows, one is able to cover large parts of images very rapidly. This random process is generic and can be applied to any kind of images. The same random process is applied to test images. Subwindows are resized to a fixed scale (16×16 pixels) and transformed to a HSV color space. Each subwindow is thus described by a feature vector of 768 numerical values. The same descriptors are used for subwindows obtained from training and test images.

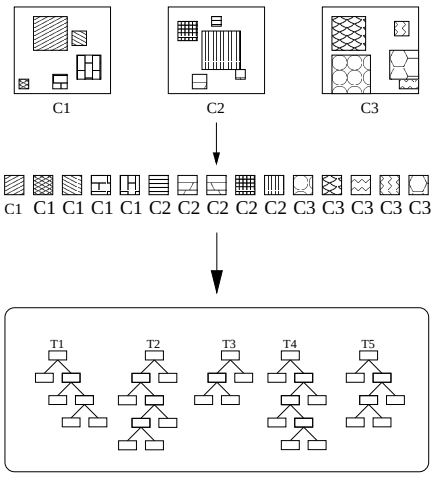


Figure 1. Learning: the framework first randomly extracts multi-scale subwindows from training-set images, then resizes them and builds an ensemble of decision trees.

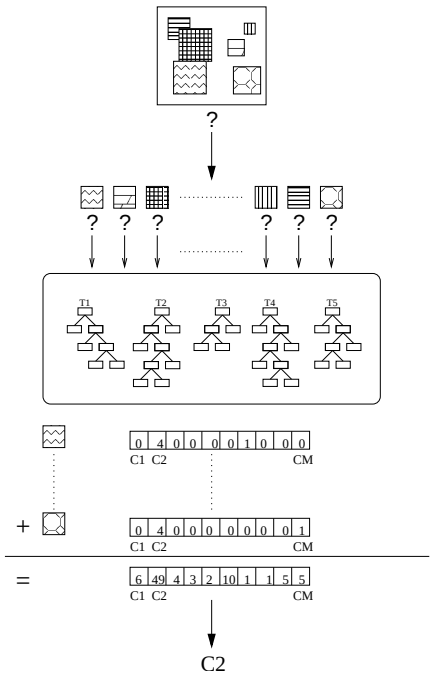


Figure 2. Recognition: randomly-extracted subwindows are propagated through the trees (here $T = 5$). Votes are aggregated and the majority class is assigned to the image.

2.2. Learning

At the learning phase, a model is automatically built using subwindows extracted from training images. First, each subwindow is labelled with the class of its parent image. Then, any supervised machine learning algorithm can be applied to build a subwindow classification model. Here, the input of a machine learning algorithm is thus a training sample of N_w subwindows, each of which is described by 768 real-valued input variables and a discrete output class (Figure 1). The learning algorithm should consequently be able to deal efficiently with a large amount of data, first in terms of the number of subwindows and classes of images in the training set, but more importantly in terms of the number of values describing these subwindows.

In this context, we compare five tree-based methods: one single-tree method based on CART (Breiman et al., 1984), and four ensemble methods: Bagging (Breiman, 1996), Boosting (Freund & Robert Schapire, 1996), Random Forests (Breiman, 2001), and Extra-Trees (Geurts, 2002). Extra-Trees only were originally used by (Marée et al., 2005).

2.3. Recognition

In this approach, the learned model is used to classify subwindows of a test image. To make a prediction for a test image with an ensemble of trees grown from subwindows, each subwindow is simply propagated into each tree of the ensemble. Each tree outputs conditional class probability estimates for each subwindow. Each subwindow thus receives T class probability estimate vectors where T denotes the number of trees in the ensemble. All the predictions are then averaged and the class corresponding to the largest aggregated probability estimate is assigned to the image. Note that we will simply consider that one single tree method is a particular case where $T = 1$.

3. Experiments

Our experiments aim at comparing decision tree methods within our random subwindow framework (Marée et al., 2005). To this end, we compare these methods on three well-known and publicly available object recognition datasets: household objects in a controlled environment (COIL-100), object categories in a controlled environment (ETH-80), and buildings in urban scenes (ZuBuD). The first dataset exhibits substantial viewpoint changes. The second dataset also exhibits higher intra-class variability. The third dataset contains images with illumination, viewpoint, scale and orientation changes as well as partial occlusions and cluttered backgrounds.

3.1. Parameters

For each problem and protocol, the parameters of the framework were fixed to $N_w = 120000$ learning sub-

windows, $T = 25$ trees, and $N_{w,test} = 100$ subwindows are randomly extracted from each test image. In (Marée et al., 2005), the parameters were fixed to $N_w = 120000$, $T = 10$, and $N_{w,test} = 100$. Ensemble methods are influenced by the number of trees T that are aggregated. Usually, the more trees are aggregated, the better the accuracy. We will further evaluate the influence of these parameters in Section 4.

For each machine learning method within the framework, the values of several parameters need to be fixed. In our experiments, single decision trees are fully developed, i.e. without using any pruning method. The score used to evaluate tests during the induction is the score proposed by (Wehenkel, 1997) which is a particular normalization of the information gain. Otherwise our algorithm is similar to the CART method (Breiman et al., 1984).

Random Forests depends on an additional parameter k which is the number of attributes randomly selected at each test node. In our experiments, its value was fixed to the default value suggested by the author of the algorithm which is the square root of the total number of attributes. According to (Breiman, 2001) this value usually gives error rates very close to the optimum.

With the latest variant of Extra-Trees (Geurts et al., 2005), the parameter k is the number of attributes randomly selected at each test node. We fixed it to the default value which is the square root of the total number of attributes. The main differences with Random Forests are that the algorithm randomizes also cut-point choice while splitting a tree node and grows the tree from the whole learning set while Random Forests uses bootstrap sampling.

Boosting does not depend on another parameter but it nevertheless requires that the learning algorithm does not give perfect models on the learning sample (so as to provide some misclassified instances). Hence, with this method, we used with decision trees the stop-splitting criterion described by (Wehenkel, 1997). It uses a hypothesis test based on the G^2 statistic to determine the significance of a test. In our experiments, we fixed the nondetection risk α to 0.005.

3.2. COIL-100

COIL-100[1] (Murase & Nayar, 1995) is a dataset of 128×128 color images of 100 different 3D objects (boxes, bottles, cups, miniature cars, etc.). Each object was placed on a motorized turntable and images were captured by a fixed camera at pose intervals of 5°, corresponding to 72 images per object. Given a new image, the goal is to identify the target object in it.

Figure 3. COIL-100: some subwindows randomly extracted from a test image and resized to 16×16 pixels.

On this dataset, reducing the number of training views increases perspective distortions between learned views and images presented during testing. In this paper, we evaluate the robustness to viewpoint changes using only one view (the pose at 0°) in the training sample while the remaining 71 views are used for testing. Using this protocol, methods in the literature yield error rates from 50.1% to 24% (Matas & Obdržálek, 2004). Our results using this protocol (100 learning images, 7100 test images) are reported in Table 1. Tree Boosting is the best method for this problem, followed by Extra-Trees, Tree Bagging, and Random Forests. One decision tree has a higher error rate. Examples of subwindows randomly extracted and resized to 16×16 pixels are given in Figure 3.

3.3. ETH-80

The Cogvis ETH-80 dataset[2] contains 3280 color images (128×128 pixels) of 8 distinct object categories (apples, pears, tomatoes, cows, dogs, horses, cups, cars). For each category, 10 different objects are provided. Each object is represented by 41 images from different viewpoints.

In our experiments, we used for each category 9 objects in the learning set ($8*9*41 = 2952$ images), and the remaining objects in the test set ($8*1*41 = 328$ images). We evaluate the methods on 10 different partitions, and the mean error rate is reported in Table 1. Here, Extra-Trees are slightly inferior while Tree Boosting and Tree Bagging are slightly better than other methods.[3]

[1] `http://www.cs.columbia.edu/CAVE/`

[2] `http://www.vision.ethz.ch/projects/categorization/eth80-db.html`

[3] We observed that the adjustment of the extra-tree parameter k to the half of the total number of attributes, instead of the square root, yields a 20.85% mean error rate. Such improvements might also be obtained for Random Forests.

Table 1. Classification error rates (in %) of all methods on COIL-100, ETH-80, ZuBuD ($T = 25$, $N_w = 120000$, $N_{w,test} = 100$)).

Methods	COIL-100	ETH-80	ZuBuD
RW+Single Tree	19.20	22.04	10.43
RW+Extra-Trees	11.53	22.74	4.35
RW+R. Forests	13.06	21.31	4.35
RW+Bagging	12.77	20.34	3.48
RW+Boosting	10.75	20.27	3.48

3.4. ZuBuD

The ZuBuD dataset[4] (Shao et al., 2003) is a database of color images of 201 buildings in Zürich. Each building in the training set is represented by five images acquired at five random arbitrary viewpoints. The training set thus includes 1005 images, while the test set contains 115 images of a subset of the 201 buildings. Images were taken by two different cameras in different seasons and under different weather conditions, and thus contain a substantial variety of illumination conditions. Partial occlusions and cluttered background are naturally present (trees, skies, cars, trams, people, ...) as well as scale and orientation changes due to the position of the photographer. Moreover, training images were captured at 640×480 while testing images are at 320×240 pixels.

About five papers have so far reported results on this dataset that vary from a 59% error rate to 0% (Matas & Obdržálek, 2004). Our results are reported in Table 1. Due to the small size of the test set, the difference between the methods is not dramatic and only one image makes the difference between the two best ensemble methods (Tree Boosting, Tree Bagging) and the two others (Extra-Trees and Random Forests). One single decision tree is again inferior. Figure 4 shows the 5 images misclassified by Extra-Trees and Random Forests, while the last one is correctly classified by Tree Boosting and Tree Bagging. For this last image, the correct class is ranked second by Extra-Trees and Random Forests.

Figure 4. ZuBuD: misclassified test images (left), training images of predicted class buildings (middle), training images of correct buildings (right).

4. Discussion

The good performance of this framework was explained by (Marée et al., 2005) by the combination of simple but well-motivated techniques: random multi-scale subwindow extraction, HSV pixel representation and recent advances in machine learning that have produced new methods that are able to handle problems of high dimensionality.

For real-world applications, it may be useful to tune the framework parameters if a specific tradeoff between accuracy, memory usage and computing times is desired. Then, in this section, we discuss the influence of the framework parameters (4.1, 4.2, 4.3) on the ZuBuD problem which exhibits real-world images, and we present some complexity results (4.4).

4.1. Variation of N_w

Figure 5 shows that the error rate decreases monotonically with number of learning subwindows (for a given number of trees ($T = 25$) and a given number of test subwindows ($N_{w,test} = 100$)). For all methods, we observe that using $N_w = 60000$ subwindows already gives good results, and that $N_w = 180000$ does not improve accuracy, except for one single decision tree.

4.2. Variation of T

Figure 6 shows that the error rate decreases monotonically with the number of trees, for a given number of training subwindows ($N_w = 120000$) and test subwindows ($N_{w,test} = 100$). We observe that using $T = 10$ trees is already sufficient for this problem for all ensemble methods.

[4] `http://www.vision.ee.ethz.ch/showroom/zubud/index.en.html`

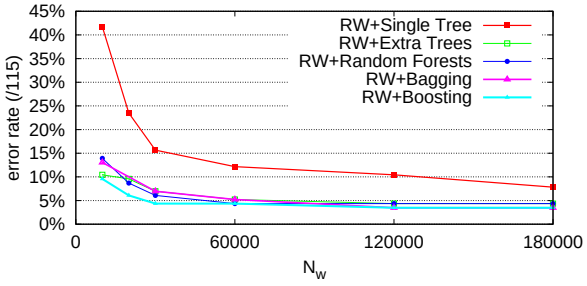


Figure 5. ZuBuD: error rate with increasing number of training subwindows.

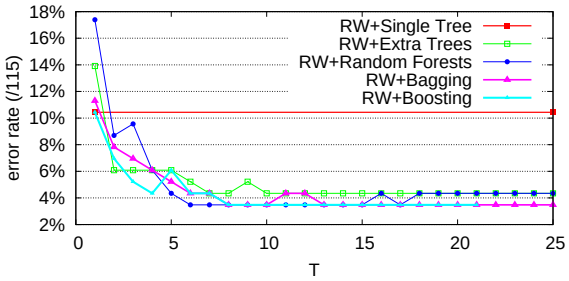


Figure 6. ZuBuD: error rate with increasing number of trees.

4.3. Variation of $N_{w,test}$

Figure 7 shows that the number of test subwindows also influences the error rate in a monotonic way, for a given number of training subwindows ($N_w = 120000$) and a given number of trees ($T = 25$). We observe that using $N_{w,test} = 25$ is already sufficient for this problem with ensemble methods, but the aggregation of more subwindows is needed for a single decision tree.

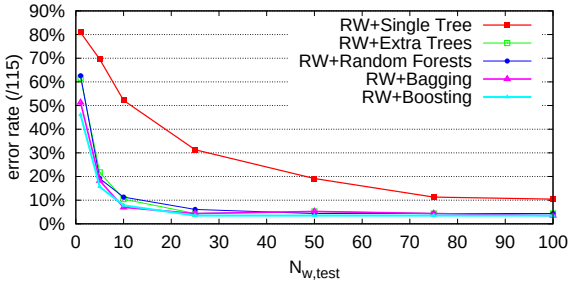


Figure 7. ZuBuD: error rate with increasing number of test subwindows.

4.4. Some notes on complexity

Our current implementation cannot be considered optimal but some indications can be given about memory and running-time requirements. With this framework, original training images and their subwindows are not necessary to classify new images after the construction of the model, contrary to classification methods based on nearest neighbors. Here, only the ensemble of trees is used for recognition.

Table 2. ZuBuD: average tree complexity and learning time.

Methods	Cmplx	Learning Time
RW+Single Tree	92687	3h36m30s
RW+Extra-Trees	148080	14m05s
RW+Random Forests	77451	2h14m54s
RW+Bagging	63285	53h35m46s
RW+Boosting	28040	54h21m31s

Learning times for one single decision tree and ensembles of $T = 25$ trees are reported in Table 2, considering that subwindows are in main memory. The complexity of tree-based method induction algorithm is of order $O(N_w log N_w)$. Extra-Trees are particularly fast due to their extreme randomization of both attributes and cut-points while splitting a tree node. Single tree complexity (number of nodes) is also given in Table 2 as a basic indication of memory usage.

To classify a new image, we observed that the prediction of one test subwindow with one tree requires on average less than 20 tests (each of which involves comparing the value of a pixel to a threshold), as reported in Table 3[5]. The minimum and maximum depths are also given. To classify one unseen image, the number of operations is thus multiplied by T, the number of trees, and by $N_{w,test}$, the number of subwindows extracted. The time to add all votes and search the maximum is negligible. Furthermore, extraction of one subwindow is very fast because of its random nature.

On this problem, we have also observed that pruning Extra-Trees could substantially reduce their complexity (downto a tree complexity average of 25191 with the same stop-splitting criterion as Tree Boosting, thus giving an average test depth of 15.4) while keeping the same accuracy. In practical applications where prediction times are essential, the use of pruning is thus certainly worth exploring.

[5]The average tree depth was calculated empirically over the 287500 propagations (100 subwindows for each of the 115 test images, propagated through $T = 25$ trees), except for one single decision tree and for Tree Boosting (because the algorithm stopped after $T = 21$ trees).

Table 3. ZuBuD: average subwindow test depth.

METHODS	DEPTH	MIN	MAX
RW+SINGLE TREE	16.59	9	29
RW+EXTRA-TREES	18.26	8	34
RW+RANDOM FORESTS	16.44	7	33
RW+BAGGING	15.98	8	34
RW+BOOSTING	15.04	6	28

5. Conclusions

In this paper, we compared 5 tree-based machine learning methods within a recent and generic framework for image classification (Marée et al., 2005). Its main steps are the random extraction of subwindows, their transformation to normalize their representation, and the supervised automatic learning of a classifier based on (ensembles of) decision tree(s) operating directly on the pixel values. We evaluated the tree-based methods on 3 publicly-available object recognition datasets. Our study shows that this general and conceptually simple framework yields good results for object recognition tasks when combined with ensembles of decision trees. Extra-Trees are particularly attractive in terms of computational efficiency during learning, and are competitive with other ensemble methods in terms of accuracy. This method with its default parameter allows to evaluate very quickly the framework on any new dataset.[6] However, if the main objective of a particular task is to obtain the best error rate whatever the learning time, Tree Boosting appears to be a better choice. Tuning the parameters (such as the value of k in Extra-Trees, or the stop-splitting criterion) might further improve the results.

For future work, it would be interesting to perform a comparative study with SVMs. The framework should also be evaluated on bigger databases in terms of the number of images and/or classes and with images that exhibit higher intra-class variability and heavily cluttered backgrounds (such as the Caltech-101[7], Birds, or Butterflies[8] datasets).

6. Acknowledgment

Raphaël Marée is supported by GIGA-Interdisciplinary Cluster for Applied Genoproteomics, hosted by the University of Liège. Pierre Geurts is a Postdoctoral Researcher at the National Fund for Scientific Research (FNRS, Belgium).

[6] Java implementation is available for evaluation at `http://www.montefiore.ulg.ac.be/~maree/`

[7] `http://www.vision.caltech.edu/feifeili/Datasets.htm`

[8] `http://www-cvr.ai.uiuc.edu/ponce_grp/data/`

References

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*, 123–140.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

Breiman, L., Friedman, J., Olsen, R., & Stone, C. (1984). *Classification and regression trees*. Wadsworth International (California).

Freund, Y., & Robert Schapire, E. (1996). Experiments with a new boosting algorithm. *Proc. Thirteenth International Conference on Machine Learning* (pp. 148–156).

Geurts, P. (2002). *Contributions to decision tree induction: bias/variance tradeoff and time series classification*. Doctoral dissertation, Department of Electrical Engineering and Computer Science, University of Liège.

Geurts, P., Ernst, D., & Wehenkel, L. (2005). Extremely randomized trees. *Submitted*.

Marée, R., Geurts, P., Piater, J., & Wehenkel, L. (2005). Random subwindows for robust image classification. *Proc. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Matas, J., & Obdržálek, S. (2004). Object recognition methods based on transformation covariant features. *Proc. 12th European Signal Processing Conference (EUSIPCO 2004)*. Vienna, Austria.

Mikolajczyk, K., & Schmid, C. (2005). A performance evaluation of local descriptors. *PAMI, to appear*.

Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., & Gool, L. V. (2005). A comparison of affine region detectors. *International Journal of Computer Vision, to appear*.

Murase, H., & Nayar, S. K. (1995). Visual learning and recognition of 3d objects from appearance. *International Journal of Computer Vision*, *14*, 5–24.

Shao, H., Svoboda, T., & Van Gool, L. (2003). *Zubud - Zurich building database for image based recognition* (Technical Report TR-260). Computer Vision Lab, Swiss Federal Institute of Technology, Switzerland.

Wehenkel, L. A. (1997). *Automatic learning techniques in power systems*. Kluwer Academic Publishers, Boston.

Multimedia Target Tracking through Feature Detection and Database Retrieval

Maria Grazia Di Bono Maria.Grazia.Dibono@isti.cnr.it
Gabriele Pieri Gabriele.Pieri@isti.cnr.it
Ovidio Salvetti Ovidio.Salvetti@isti.cnr.it
Institute of Information Science and Technologies, ISTI-CNR, Via Moruzzi 1, 56124 Pisa, ITALY

Abstract

The real-time detection and tracking of moving objects is a challenging task and automatic tools to identify and follow them are often subject to constraints regarding the environment under investigation or the full visibility of the targeted object. Exploiting the possibility of a multi-source acquisition in the targeted scene, firstly detection is performed by means of characteristic features extraction and storing in a database; secondly, the tracking task is approached using algorithms, where automatic search involves occluded or masked targets in the scene. This latter problem is solved through database retrieval, based on well-defined multi-modal features. The method has been tested on case studies regarding the identification and tracking of animals moving at night in an open environment (i.e. natural reserves or parks), and the surveillance of known scenes for unauthorized access control.

1. Introduction

According to the cognitive processes of the human perception (Milner & Goodale, 1995), a methodology has been developed which provides a way to realize object recognition and tracking in 3D real environments. In particular, this approach is based on the acquisition of multi-source information that is firstly elaborated for object detection and characterization, and then for its localization and active tracking. After target detection is achieved through an automatic segmentation, the characterization phase is performed through the description of multi-modal features (morphological, densitometric and semantic), which are extracted from the acquired multi-source information. Localization is realized using also features previously extracted and stored in a reference database. In order to improve the localization performance when only partial information is available (i.e. in case of lost or occluded targets), the implemented method is supported by a content-based retrieval (CBR) paradigm using an a priori defined multimedia (MM) database. This MM database is built using the multi-modal features extracted from a set of target examples organised on the basis of semantic classes defined on the specific environment under investigation.

Current approaches regarding real-time object tracking from videos are based on (i) successive frame differences (Fernandez-Caballero et al., 2003), using also adaptive threshold techniques (Fejes & Davis, 1999), (ii) trajectory tracking, using weak perspective and optical flow (Yau, Fu & Liu, 2001), (iii) region approaches, using active contours of the target and neural networks for movement analysis (Tabb et al., 2002), or motion detection and successive regions segmentation (Kim & Kim, 2003).

Regarding the CBR paradigm, techniques of shape retrieval in large databases are particularly interesting. Considering a shape of an object as a sequence of contour points, a method using both global and local features is discussed in (Wang, Yang & Acharya, 1998) while in (Wang, Chang & Acharya, 1999) retrieval is based on a hash table and a majority voting algorithm for an efficient estimation of shape similarity. Furthermore, another interesting approach considers a shape database structured as an *M-tree* of organised tokens, representing parts of the shape enclosed between contour points. Possible shapes are clustered into semantic classes, each belonging to an object typology defined in its environment (Berretti, Del Bimbo & Pala, 2000).

In this paper, the problem of moving target detection and tracking is faced by processing multi-source information acquired using cameras of different typology (Far-IR and visible). Object characterization is based on region segmentation and feature extraction processes. Object localization uses a CBR approach based on similarity functions defined for each multi-modal feature class.

The method has been applied to real case studies regarding the monitoring of animal movements during the night in an open environment (i.e. natural reserves or parks) and the surveillance of known scenes for

Appearing in *Proceedings of the workshop Machine Learning Techniques for Processing Multimedia Content*, Bonn, Germany, 2005.

unauthorized access control in both open and closed spaces (Pieri et al., 2004).

2. Problem definition

The precise identification of a defined target in a real video, frame by frame, is approached. The proposed methodology is based on *recognition* and *spatial localization* of the target: recognition is sub-divided into identification and characterization, while the spatial localization performs active tracking.

The multi-source information is acquired using a physical system composed of a thermo-camera and two stereo visible-cameras synchronized. Thus, we obtain a set of infrared (IR) images, which make the system more robust and invariant to light changes in the scene, corresponding to stereo grey level images.

A procedure has been defined based on two different stages:

- *Off-line stage*, in which the recognition phase is performed using selected examples belonging to a set of predefined semantic classes, in order to populate the reference MM database.
- *On-line stage*, in which the tracking is performed by applying recognition and spatial localization.

In deep details, during the recognition process, the identification phase consists of an automatic segmentation, based on edge detection using a gradient descent along 16 directions starting from a reference point internal to the target (centroid).

In the characterization phase, for each frame, the multi-source information is used in order to extract a target description from the scene. This is made through a feature extraction process performed on the three different images available for each frame in the sequence. In particular, the extraction of a depth index from the grey level stereo images, performed by computing disparity of the corresponding stereo points, is realized in order to have significant information about the target spatial localization in the 3D scene and the target movement along depth direction, which is useful for the determination of a possible static or dynamic occlusion of the target itself in the observed scene. Other features consisting in radiometric parameters measuring the temperature and visual features are extracted from the IR images. The visual features, grouped in *morphological*, *densitometric* and *semantic* classes, consist of shape contour descriptors, dominant colour discriminants, statistical parameters, computed on the regions enclosed by the contours (area, perimeter, average brightness, standard deviation, skewness, kurtosis, and entropy) and the semantic class to which the target belongs (i.e. human, small, medium and large animal, …). While the depth index and the visual features are automatically extracted from the images, the semantic classes of the observed targets are selected by the user among a predefined set of possible choices.

During the off-line stage, all the multi-modal feature information is stored in the MM database, organised on the base of semantic classes. This information is used during the on-line spatial localization process, in particular in the automatic target retrieval which acts as a support during the active tracking in case of partial occlusion or quickly direction changes of the target.

For each defined target class, possible variations of the initial shape, taking into account that the target could be still partially masked or have a different orientation, are recorded together with the other multi-modal features as it is shown in Figure 1.

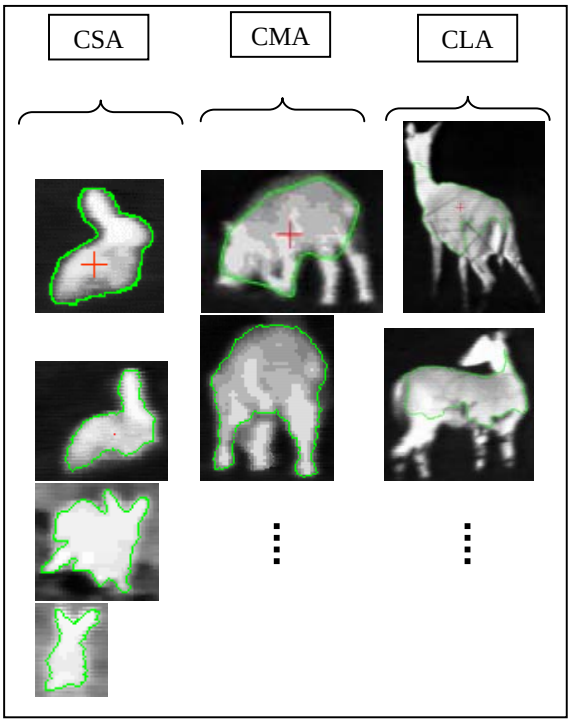


Figure 1. Example of actual targets in the MM data-base, grouped according to the classes “small animal” (CSA), “medium animal” (CMA) and “large animal” (CLA).

During the on-line spatial localization, the extracted features drive tracking and also support CBR to resolve the queries to the MM database.

The first phase of the on-line stage is the same of the off-line one. An automatic segmentation of the target is performed on both the IR and stereo grey-level images, in order to characterize the selected target. Contextually, the user performs also the selection of the semantic class to which the target belongs.

The tracking algorithm is performed on the IR image sequence, in order to build a system which can be used both at night and daylight.

The segmented target from the first IR image of the sequence is then tracked automatically in the following frame. The features used for the automatic tracking are *local maxima*, *movement prediction* (on the basis of the movements of the previous steps), *temperature* and *a priori knowledge* about the specific class the object belongs to. For each frame, the algorithm performs the steps to correctly identify the target and to follow it.

Firstly, a candidate characterizing point P_1 of the target is selected in its *centroid*, in the actual frame. The selection follows criteria of brightness local maximum, inside the contour segmented in the previous frame; P_1 is the point having the maximum similarity with the centroid P_P of the previous frame.

In a second step, the algorithm takes into account the previous movements of the centroid. The trajectory is stored and then used in the computation of the actual step, locating a new candidate point P_2. If P_2 is not coincident with P_1 then a new point P_3 is calculated as:

$$P_3 = \alpha P_1 + \beta P_2 \qquad (1)$$

where α and β represent the weight assigned, and $\alpha + \beta = 1$. These parameters are empirically defined and can be adjusted by the user.

Again, a local maximum search is performed in the neighbourhood of P_3 to make sure that it is internal to a valid object. This search finds the point P_N that has the grey level closest to the one of P_P, so that P_N is the centroid chosen for the actual frame. Starting from this point, the edge detection is performed and the object new contour is segmented.

In each frame, a first control is made trying to avoid a wrong object recognition, due to either a masking, partial occlusion of the object in the scene or to a quick movement in an unexpected direction. This control takes into account the above mentioned statistical parameters computed on the region enclosed by the contour, without using CBR paradigm in order to optimise the number of accesses to the database. If there are parameters exceeding *p* times (*p* is defined a priori) the standard deviation of the same parameters computed over the last *n* frames, the database search for the correct target is started. This search is based on the CBR paradigm; the multi-modal features of the candidate target are compared to the ones recorded in the MM database. A similarity function is considered for each feature class. In particular, we used similarity functions, as in (Tzouveli et al., 2004), for *colour matching*, using percentages and colour values, and *shape matching*, using the cross-correlation criterion. In order to obtain a global similarity measure, each similarity percentage is associated to a pre-selected weight, using the reference semantic class as a filter to access the MM information. If after *j* frames the correct target has not yet been grabbed, the control is given back to the user. The value of *j* is computed considering the distance between P_P and the edge point of the image along the search direction, divided by the average velocity of the target previously measured in the last *n* frames (Eq. 2).

$$j = \mathrm{Dist}(P_P; E_r) / \mathrm{Vel} \qquad (2)$$

where $\mathrm{Dist}(x, y)$ is the Euclidean distance between points *x* and *y*; E_r is the point crossing the edge of the frame along the search direction *r* determined by the last *n* centroids; and *Vel* is

$$\mathrm{Vel} = \left(\sum_{i=0}^{n-1} \mathrm{Dist}(P_P^i; P_P^{i+1}) \right) / n \qquad (3)$$

where P_P^i is the centroid *i* steps before the actual.

The sketch of the methodology described is shown in Figure 2.

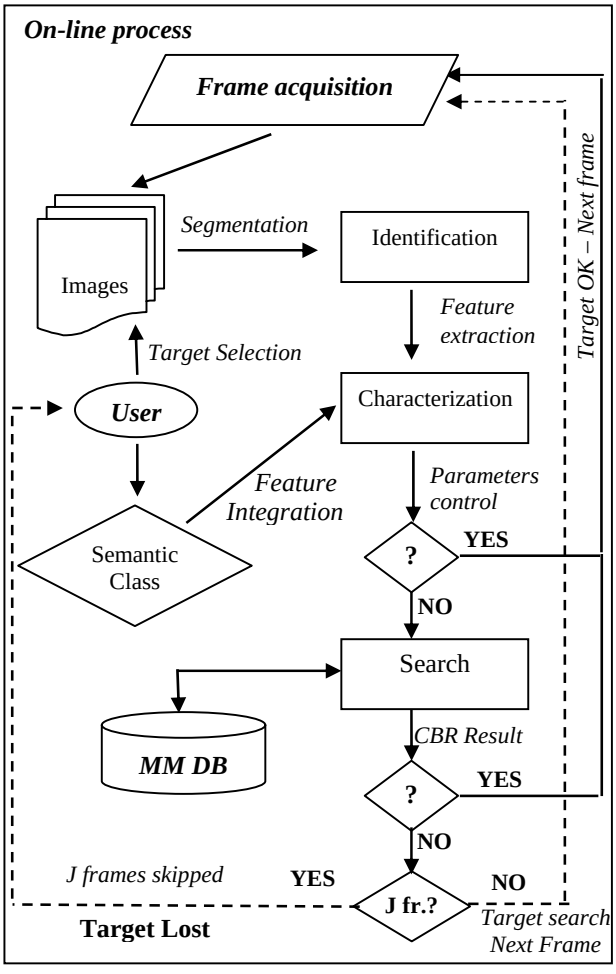


Figure 2. Recognition and description of a target object (on-line process).

3. Results and Conclusions

The method implemented has been applied to real case studies: (i) to track animal movements in an open environment during the night, for the fauna monitoring in natural parks, and (ii) for video surveillance of known scenes both at night and daylight to control unauthorized access (see Figure 3).

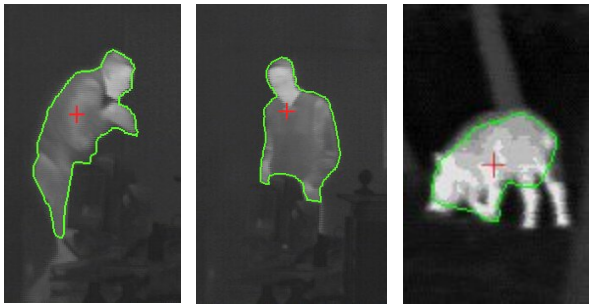


Figure 3. Examples of thermo images regarding human (left and centre, two different views and shapes) and animal (right) targets (crosses are the centroids).

Regarding the first case, due to the environmental conditions, only the thermo-camera has been used.

The videos were acquired using a thermo-camera in the 8-12μm wavelength range, mounted on a moving structure covering 360° pan and 90° tilt, and equipped with 12° and 24° optics to have 320x240 pixel spatial resolution.

Both the thermo-camera and the two stereo visible-cameras have been positioned in order to explore a scene 100 meters far, sufficient in our experimental cases.

In the fauna monitoring experimental case, during the off-line stage, the MM database has been built taking into account different image sequences relative to different classes of the monitored animals. In particular, three main semantic classes have been determined. The *large-animal* class counting all the monitored animals of a large size like deer, the *medium-animal* class including animals of medium size like boars and the *small-animal* class considering other kind of animals like rabbits or badgers. For each outlined semantic class, different positions have been considered. In more details, four different positions for boars, rabbits and other small animals and six for deer have been registered.

In the video-surveillance case, the *human* class has been composed taking into account six different pose conditions for three different people typology.

The acquired images are pre-processed to reduce the noise, the algorithm has shown an effective performance and seems promising in the lights of further improvements regarding for example the integration with *audio information*, coming from different aligned microphones installed in the scene, and aiming at the same direction of the cameras.

References

Berretti, S., Del Bimbo, A., & Pala, P. (2000). Retrieval by Shape Similarity with Perceptual Distance and Effective Indexing. *IEEE Transactions on Multimedia, 2 (4)*, 225–239.

Fejes, S., & Davis, L.S. (1999). Detection of Independent Motion Using Directional Motion Estimation. *Computer Vision and Image Understanding, 74 (2)*, 101–120.

Fernandez-Caballero, A., Mira, J., Fernandez, M.A., & Delgado, A.E. (2003). On motion detection through a multi-layer neural network architecture. *Neural Networks, 16*, 205–222.

Kim, J.B., & Kim, H.J. (2003). Efficient region-based motion segmentation for a video monitoring system. *Pattern Recognition Letters, 24*, 113–128.

Milner, A.D., & Goodale, M.A. (1995). *The Visual Brain in Action*. Oxford: Oxford University Press.

Pieri, G., Benvenuti, M., Carnier, E., & Salvetti, O. (2004). Object detection and tracking in an open and free environment with a moving camera. *Proceedings of Seventh International Conference on Pattern Recognition and Image Analysis: New Information Technologies* (Vol. II, pp. 347–350), St. Petersburg, Russia, 18-23 October 2004.

Tabb, K., Davey, N., Adams, R., & George, S. (2002). The recognition and analysis of animate objects using neural networks and active contour models. *Neurocomputing, 43*, 145–172.

Tzouveli, P., Andreou, G., Tsechpenakis, G., Avrithis, Y., & Kollias, S. (2004). Intelligent Visual Descriptor Extraction from Video Sequences. *Lecture Notes in Computer Science – Adaptive Multimedia Retrieval, 3094*, 132–146.

Wang, J., Yang, W.-J., & Acharya, R. (1998). Efficient Access to and Retrieval from a Shape Image Database. *Proceedings of the IEEE Workshop on Content-Based Access of Image and Video Libraries* (pp. 63–67). Santa Barbara, CA.

Wang, J., Chang, W., & Acharya, R. (1999). Efficient and Effective Similar Shape Retrieval. *Proceedings of the IEEE International Conference on Multimedia Computing and Systems* (Vol. 1, pp. 875–879), Florence, Italy.

Yau, W.G., Fu, L.-C., & Liu, D. (2001). Robust Real-time 3D Trajectory Tracking Algorithms for Visual Tracking Using Weak Perspective Projection. *Proceedings of the American Control Conference*, Arlington, VA.

Active Learning Techniques for User Interactive Systems: Application to Image Retrieval

Philippe Henri Gosselin GOSSELIN@ENSEA.FR
Matthieu Cord CORD@ENSEA.FR
ETIS / CNRS UMR 8051, 6 avenue du Ponceau, 95014 Cergy-Pontoise, France

Abstract

Active learning methods have been considered with an increasing interest for user interactive systems. In this paper, we propose an efficient active learning scheme to deal with this particular context. An active boundary correction is proposed in order to deal with few training data. Experiments are carried out on the COREL photo database.

1. Introduction

Human interactive systems has attracted a lot of research interest in recent years, especially for content-based image retrieval systems. Contrary to the early systems, focused on fully automatic strategies, recent approaches introduce human-computer interaction (Veltkamp, 2002; Vasconcelos & Kunt, 2001).

Starting with a coarse query, the interactive process allows the user to refine his request as much as necessary. Many kinds of interaction between the user and the system have been proposed (Chang et al., 2003), but most of the time, user information consists of binary annotations (labels) indicating whether or not the image belongs to the desired category.

In this paper, we focus on the retrieval of *concepts* within a large document collection. We assume that a user is looking for a set of documents, the query concept, within an existing document database. The aim is to build a fast and efficient strategy to retrieve the query concept.

Performing an estimation of the query concept can be seen as a statistical learning problem, and more precisely as a binary classification task between the relevant and irrelevant classes (Chapelle et al., 1999). The relevant class is the set of documents within the query concept, and the irrelevant class the set of documents out of the query concept. This context defines a very specific learning problem with the following characteristics:

1. *High dimensionality.* The documents used to be represented by vectors of high dimensionality.
2. *Few training data.* At the beginning, the system has to perform a good estimation of the query concept with very few data. Furthermore, the system can not ask user to label thousands of documents, good performances are required using a small percentage of labeled data.
3. *Relevance feedback.* Due to user annotations, the training data set grows step by step during the retrieval session, so the current classification depends on the previous ones.
4. *Unbalanced classes.* The query concept is often a small subset of the database (some hundreds of documents). Thus, the relevant and irrelevant classes are highly unbalanced (up to factor 100), on the contrary to classical classification problems, where the classes have approximatively the same size.
5. *Limited computation time.* The user can not wait several hours between each feedback steps. We assume that a user can wait at most several minutes between each feedback steps.

In this paper, we propose an active learning strategy to deal with these characteristics. In section 2, we present current methods for classification, and motivations for active learning. In section 3, we focus on active learning, and present two well-known approaches: uncertainly-based sampling and error reduction. In section 4, we propose an active learning scheme to enhance the previous methods. In section 5, experiments

Appearing in *Proceedings of the workshop Machine Learning Techniques for Processing Multimedia Content*, Bonn, Germany, 2005.

are carried out on a generalist image database in order to compare the different strategies.

2. Learning for human interactive systems

2.1. Kernels and SVM

The first characteristic to deal with is the high dimensionality of feature vectors. With vectors of high dimensionality (for instance, 100 or more), artifacts appear, known as the result of the curse of dimensionality (Hastie et al., 2001). However, with the theory of kernel functions, one can reduce this curse (Smola & Scholkopf, 2002), especially if one can build a kernel function for a specific application. For instance, when distributions are used as feature vectors, a Gaussian kernel gives excellent results in comparison to distance-based techniques (Gosselin & Cord, 2004a).

Using a kernel function leads to a set of classification methods. For human interactive systems, statistical learning techniques such as nearest neighbors (Hastie et al., 2001), support vector machines (Tong & Chang, 2001; Chapelle et al., 1999; Chen et al., 2001), bayes classifiers (Vasconcelos & Kunt, 2001), have been used. We have previously shown that the SVM classification method is highly adapted to the image retrieval context (Gosselin & Cord, 2004a). Thus, we will use SVM as classification method in the following sections.

2.2. Semi-supervised learning

A natural choice for dealing with the second characteristic – the few training data – is to use semi-supervised learning techniques. Semi-supervised techniques uses labeled and unlabeled documents to compute a classification function. For instance Transductive SVM (Joachims, 1999), semi-supervised Gaussian mixtures (Najjar et al., 2003), and semi-supervised Gaussian fields (Zhu et al., 2003). However, TSVM and SSGM do not lead to significant improvements (Chang et al., 2003; Gosselin et al., 2004). Furthermore, these techniques have high computational needs in comparison to inductive techniques, and sometimes untractable. For instance, SSGF needs the inversion of a $N \times N$ matrix, where N is the size of the database. For now, semi-supervised learning techniques do not seem to be adapted to the context we are focusing on.

2.3. Active learning

Active learning is another solution to deal with few training data. The interaction between the user and the system can be exploited. The user is able to label any document in the database. The only constraint is that this user will not label a lot of documents. However, even a small labelling lead to significant improvements with active learning.

3. Active learning strategies

In this paper, we focus on the active learning scheme where a pool unlabeled examples is available. We suppose that we have a set $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ of documents, a set of labels $\mathbf{y} = (y_1, \ldots, y_N)$ (1 relevant, -1 irrelevant, 0 unknown), a relevance function $f_{\mathbf{y}} : \mathbf{X} \rightarrow [-1, 1]$ trained with $\mathbf{y}$, and a teacher $\tau : \mathbf{X} \rightarrow \{-1, 1\}$ that labels documents as -1 or 1. We also denote by I the set of indexes of labeled documents.

The aim of an active learning within this context is to choose the unlabeled document $\mathbf{x}$ that will enhance the most the relevance function trained with the label $\tau(\mathbf{x})$ added to the previous labeling $\mathbf{y}$. We propose to formalize this choice as the minimization of a cost function $g(\mathbf{x})$ over all unlabeled documents. Thus, according to a particular active learning method, the chosen document to label is the argument of the minimum of $g(\mathbf{x})$. We also denote by J the set of candidates, *i.e.* the indexes of unlabeled documents evaluated by $g(\mathbf{x})$.

We present here two active learning strategies: uncertainly-based sampling, which selects the documents for which the relevance function is most uncertain about, and error reduction, which aims at minimizing the generalization error of the classifier. We also present a strategy for batch selection.

3.1. Uncertainly-based sampling

This strategy aims at selecting unlabeled documents that the learner of the relevance function is most uncertain about. The first solution is to compute a probabilistic output for each documents, and select the unlabeled documents with the probabilities closest to 0.5 (Lewis & Catlett, 1994). Similar strategies have been also proposed with SVM classifier (Park, 2000), with a theoretical justification (Tong & Koller, 2001), and with nearest neighbor classifier (Lindenbaum et al., 2004).

In all cases, a relevance function may be computed. This function can be a distribution, a fellowship to a class (distance to the hyperplane for SVM), or a utility function. Thus, with some adaptation of each approach, a relevance function $f_{\mathbf{y}} : \mathbf{X} \rightarrow [-1, 1]$ is trained, where the most uncertain documents have an output close to 0. The cost function to minimize is then $g(\mathbf{x}) = |f(\mathbf{x})|$.

With such a strategy, the efficiency of a method depends on the accuracy of the relevance function estimation close to 0. This is the area where it is the most difficult to perform a good evaluation[1]. In this particular context, statistical techniques are not always the best ones, and we propose in the next section an heuristic-based correction to the estimation of $f_{\mathbf{y}}$ close to 0.

3.2. Error Reduction

Active learning strategies based on error reduction select documents that, once added to the training set, minimize the error of generalization (Roy & McCallum, 2001).

Let $P(c|\mathbf{x})$ the (unknown) probability of a document $\mathbf{x}$ to be in class c, and $P(\mathbf{x})$ the (also unknown) distribution of the documents. A training set $\mathcal{A}$ with pairs $(\mathbf{x}, c)$ sampled from $P(\mathbf{x}), P(c|\mathbf{x})$ provides the estimation $\hat{P}_{\mathcal{A}}(c|\mathbf{x})$ of $P(c|\mathbf{x})$. The expected error of generalization can be written as:

$$E_{\hat{P}_{\mathcal{A}}} = \int_{\mathbf{x}} L(P(c|\mathbf{x}), \hat{P}_{\mathcal{A}}(c|\mathbf{x}))dP(x)$$

with L a loss function which evaluates the loss between the estimation $\hat{P}_{\mathcal{A}}(c|\mathbf{x})$ and the true distribution $P(c|\mathbf{x})$.

The optimal pair $(\mathbf{x}^\star, c^\star)$ is the one which minimizes this expectation:

$$\forall(\mathbf{x}, c) \quad E_{\hat{P}_{\mathcal{A}^\star}} < E_{\hat{P}_{\mathcal{A}+(\mathbf{x},c)}}$$

with $\mathcal{A}^\star = \mathcal{A} + (\mathbf{x}^\star, c^\star)$.

Roy and McCallum propose to estimate the probability $P(c|\mathbf{x})$ with the relevance function provided by the classifier, and estimate $P(\mathbf{x})$ over $\mathbf{X}$. With a maximum loss function, the estimation of the expectation becomes, with J the set of unlabeled documents:

$$\hat{E}_{\hat{P}_{\mathcal{A}^\star}} = \frac{1}{|J|} \sum_{\mathbf{x} \in J} \left(1 - \max_{c \in \{-1,1\}} \hat{P}_{\mathcal{A}^\star}(c|\mathbf{x})\right)$$

As We don't know the label of each candidate. Roy and McCallum compute the expectation for each possible label, which finally gives the following cost function:

$$g(\mathbf{x}) = \sum_{c \in \{-1,1\}} E_{\hat{P}_{\mathcal{A}+(\mathbf{x},c)}} \hat{P}_{\mathcal{A}}(c|\mathbf{x})$$

with $\hat{P}_{\mathcal{A}}(c|\mathbf{x})$ estimated with the relevance function $f_{\mathbf{y}}(\mathbf{x})$:

$$\hat{P}_{\mathcal{A}}(c|\mathbf{x}) = \frac{c}{2}(f_{\mathbf{y}}(\mathbf{x}) + c)$$

with $f_{\mathbf{y}}(\mathbf{x})$ such as $\mathbf{y}$ encodes the training set $\mathcal{A}$.

[1] In the context of human interactive system, where only few training data is available, this is a major problem.

3.3. Batch selection

In human interactive systems, it is often necessary to select batches of new training examples. A lot of active learning strategies are made to select only one new training example. With no particular extension, these strategies can select several instances very close in the feature space. Considering the power of current classification techniques, labeling a batch of very close documents or only one of them always gives the same classification.

In the version space reduction scheme, (Tong & Koller, 2000) propose to select batches yielding minimum worst-case version space volume. However, this method requires a lot of computations making it infeasible in practice. (Brinker, 2003) proposes a fast approximation of this strategy, based on the diversity of angles between the hyperplanes in the version space. The method selects documents close to the SVM boundary one far from another, and also far from the current training data:

$I^\star = 0$
repeat
  $t = \underset{i \in J}{\operatorname{argmin}} \; (\lambda |f(\mathbf{x}_i)| + (1-\lambda) \max_{j \in I \cup I^\star} k^\star(\mathbf{x}_i, \mathbf{x}_j))$
  $I^\star = I^\star \cup \{\mathbf{x}_t\}$
until $|I^\star| = l$

with $\lambda \in [0, 1]$ and $k^\star(\mathbf{x}_i, \mathbf{x}_j)$ the angle between two instances:

$$k^\star(\mathbf{x}_i, \mathbf{x}_j) = \frac{|k(\mathbf{x}_i, \mathbf{x}_j)|}{\sqrt{k(\mathbf{x}_i, \mathbf{x}_i) k(\mathbf{x}_j, \mathbf{x}_j)}}$$

The λ parameter can be used to adjust the diversity strategy contribution; $\frac{1}{2}$ is chosen as default value [2].

4. Active learning scheme

For both active learning strategies, the estimation of the relevance function is decisive. We propose in the following subsection an active correction to deal with very few training data (less than 1%). We also propose an active learning scheme with diversity for any cost function-based active learning method, and a practical solution to reduce the computation time.

4.1. Active Boundary Correction

We propose to perform the following correction to the relevance function:

$$f^\star(\mathbf{x}) = f(\mathbf{x}) - f(\mathbf{x}_{O_s})$$

[2] If additional knowledge is available (for instance, keywords), it can be used to tune this parameter.

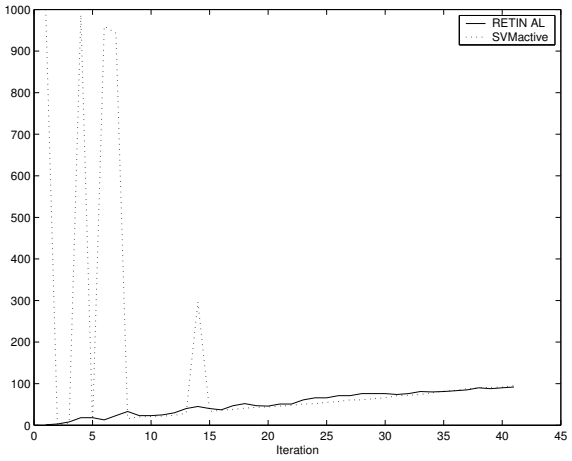


Figure 1. Values of s(t) according to feedback steps.

where $O = \text{argsort } f$, and s the correction index.

The aim of this approach is to compute s such as the "ideal" relevance function is zero at $\mathbf{x}_{O_s}$. To perform this, we propose to use the interaction with the user. The idea is to ensure that the user labels as many relevant as irrelevant documents. Then, the selected area is the most uncertain one. If the user provides a lot of relevant labels, we assume that we are close to the heart of the relevant class. Then we move the selected area further from the heart of the relevant class. If the user provides a lot of irrelevant labels, we assume that we are far from the relevant class, and then move the selected area closer to the heart of the relevant class.

We define a document $\mathbf{x}$ as close to the heart of the relevant class as $f(\mathbf{x})$ is close to 1. We change the correction index s after a sort O of the documents according to the relevance function $f(\mathbf{x})$. Small values of s means that the zero of the ideal relevance function is close to the heart of the relevant class, and vice-versa. At the feedback iteration t, we assume that the "ideal" relevance function is zero at $\mathbf{x}_{O_{s(t)}}$. We compute the new correction index $s(t+1)$ according to the labels given by the user:

$$s(t+1) = s(t) + h(\text{pos}(t), \text{neg}(t))$$

with $\text{pos}(t)$ (resp. $\text{neg}(t)$) is the number of relevant (resp. irrelevant) labels provided by the user at the feedback iteration t, and $h(a,b)$ an heuristic function. In order to get the desired behavior, we propose the following heuristic function: $h(a,b) = 2 \times (a-b)$.

At step $t = 0$, because we have no idea of the level of complexity of the searched concept, we set $s(0) = 0$.

This method is especially interesting in a context with training data, where the estimation of $f(\mathbf{x})$ is difficult. We compared this method with SVM_{active} on an image database, with 5 labels per iteration (see Experiments Section for further details). The curves in Figure 1 shows the values of $s(t)$ according to feedback steps. For the SVM_{active} method, $g(\mathbf{x}) = |f(\mathbf{x})|$ and s is such as $f(\mathbf{x}_{O_s})$ is closest to 0. Both methods have the same behavior, but SVM_{active} is very unstable during the first iterations.

This correction can be used with the active learning methods presented in the previous section. For uncertainly-based, this is simple. For error reduction, the correction is applied each time a classifier is computed. The same correction is applied, according to a new ranking of the new relevance function. A new value $f_{\mathbf{y}}(\mathbf{x}_{O_s})$ is computed for each new training set $\mathbf{y}$, and each case the relevance function is such as $f^{\star}_{\mathbf{y}}(\mathbf{x}_{O_s}) = 0$.

4.2. Incorporating diversity

In order to select batches with diversity, we propose to use the angle diversity scheme (with $g(\mathbf{x})$ instead of $|f(\mathbf{x})|$):

$I^{\star} = 0$
repeat
  $t = \underset{i \in J}{\text{argmin}} \; (\lambda g(\mathbf{x}_i) + (1-\lambda) \max_{j \in I \cup I^{\star}} k^{\star}(\mathbf{x}_i, \mathbf{x}_j))$
  $I^{\star} = I^{\star} \cup \{\mathbf{x}_t\}$
until $|I^{\star}| = l$

We normalize the cost function $g(\mathbf{x})$ before performing this step, in order to get values in the same interval than the cosines value interval. We observe that a diversity technique allows to select documents for labeling which are not close one to another. It is decisive in image retrieval context.

4.3. Reduce the computation time

In order to propose labels to a human expert in a reasonable time, all unlabelled documents can not be evaluated. We propose to restrict the evaluation of $g(\mathbf{x})$ to a set of *candidates*. We denote by J the set of the indexes of these candidates. We propose to reduce the set of unlabeled documents to the m closest documents to the boundary. For methods using boundary correction, the correction is made before the selection of the candidates. Thus, the boundary correction also changes the choice of the candidates.

5. Experiments

5.1. Evaluation Protocol

Tests are carried out on the generalist COREL photo database, which contains more than 50,000 pictures. To get tractable computation for the statistical evaluation, we randomly selected 77 of the COREL folders, to obtain a database of 6,000 images. To perform interesting evaluation, we built from this database 50 concepts[3]. Each concept is built from 2 or 3 of the COREL folders. The concept sizes are from 50 to 300. The set of all the concepts covers the whole database, and each image of the database is at least in one of the concepts, and at most in 5 different concepts.

Color and texture distributions are used as feature vectors, the kernel is a Gaussian kernel with a χ^2 distance, and the classification method is SVM.

We simulate the use of a image retrieval system. For one retrieval session, we assume that the user chooses one picture in the database for the concept he is looking for. A concept and a picture from this concept are randomly chosen for each new simulated retrieval session. In practice, this is done when the user brings one picture of its own. Then, the system computes features of this picture, and labels as relevant the closest picture. Other techniques could be used for this, for instance using keywords.

Thus, the simulated retrieval session starts with one relevant picture. Next, the system asks the active learner for 5 images to label. These images are labeled according to the desired concept, and the system asks again the active learning for 5 other images to label, using the 6 current labels. These feedback steps are iterated 10 times, and at the end of the retrieval session, the training set has 51 labels. Using these labels, a classification of the database is performed. The error of classification and the number of pictures in the concept within the 100 most relevant ones (top-100) are computed. We simulate 1,000 retrieval sessions for each active learning method. The error of classification and the top-100 are averaged over all retrieval sessions.

5.2. Comparison

Results are reported in Figure 2 with a full set of candidates (all unlabeled documents), and in Figure 3 with a reduced set of 100 candidates. The first line shows the active learning method. The "None" method means that no classification is performed, only the distance between an image and all other ones is computed. The second line shows the Top-100 for each method. For the "None" method, this result means that the average probability to find an image within the same concept than the considered image in the 100 nearest neighbors is 16%. The third line shows the average classification error. The last line shows the average computation time for a retrieval session.

The error reduction method (ER) gives better results than the uncertainly-based method (UB) (*cf.* Fig. 2). However, much more computation time is required (*cf.* Fig. 2) for ER, and it does not well support the reduction of the set of candidates in terms of classification error (*cf.* Fig. 3). The angle diversity improvement (AD) increases performances in all cases. This shows that, even if this method was built especially for UB, it can be used with others active learning methods. Furthermore, its costs in terms of computation time is small. The active boundary correction (BC) also increases performances in all cases. It has also a negligible cost in terms of computation time, and well supports the reduction of the set of candidates. Note that the improvement is much more significant for UB than for ER. Globally, the reduction of the set of candidates is interesting for all strategies, expect for ER without BC. For comparable performances, the computation time is divided by 10. Finally, the most efficient strategy is the BC+UB+AD strategy, which combines boundary correction, uncertainly-based, and angle diversity.

6. Conclusion

In this paper, we proposed active learning strategies for interactive search systems. We introduced an algorithm to correct the boundary of a classifier function, in order to improve the active learning efficiency. We proposed an active learning scheme combining different techniques, and a method to reduce the computation time. These strategies have been compared on a generalist image database. Results show that the efficiency of the proposed combinations, especially our strategy using boundary correction, uncertainly-based, and angle diversity. These results also show that the computation time can be significantly reduced using the proposed method without dramatical degradation of performances.

[3] A description of this database and the 50 concepts can be found at: http://www-etis.ensea.fr/∼cord/data/mcorel50.tar.gz. This archive contains lists of image file names for all the concepts.

Method	None	UB	ER	UB+AD	ER+AD	BC+UB+AD	BC+ER+AD
Top-100	16	28	33	31	34	**36**	35
Classification Error	–	8.2%	6.7%	6.7%	4.3%	**2.5%**	3.0%
Time	0.07s	0.41s	600s	60s	700s	60s	700s

Figure 2. Average Top-100, classification error and execution time for each active learning method, 10 feedbacks steps, 5 labels per step, with full set of candidates (all unlabeled documents). Legend: UB = Uncertainly-Based, ER = Error Reduction, AD = Angle diversity, BC = Boundary Correction.

Method	None	UB	ER	UB+AD	ER+AD	BC+UB+AD	BC+ER+AD
Top-100	16	28	32	31	34	**36**	35
Classification Error	–	8.4%	19.3%	6.9%	13.7%	**2.6%**	3.1%
Time	0.07s	0.41s	5.4s	2.4s	6.8s	2.4s	6.8s

Figure 3. Same protocol as Fig.2 with a reduced set of 100 candidates.

References

Brinker, K. (2003). Incorporating diversity in active learning with support vector machines. *International Confenrence on Machine Learning* (pp. 59–66).

Chang, E., Li, B. T., Wu, G., & Goh, K. (2003). Statistical learning for effective visual information retrieval. *IEEE International Conference on Image Processing*. Barcelona, Spain.

Chapelle, O., Haffner, P., & Vapnik, V. (1999). Svms for histogram based image classification. *IEEE Transactions on Neural Networks*, *9*.

Chen, Y., Zhou, X., & Huang, T. (2001). One-class svm for learning in image retrieval. *International Conference in Image Processing (ICIP'01)* (pp. 34–37). Thessaloniki, Greece.

Gosselin, P., & Cord, M. (2004a). A comparison of active classification methods for content-based image retrieval. *International Workshop on Computer Vision meets Databases (CVDB), ACM Sigmod*. Paris, France.

Gosselin, P., & Cord, M. (2004b). RETIN AL: An active learning strategy for image category retrieval. *IEEE International Conference on Image Processing*. Singapore.

Gosselin, P., Najjar, M., Cord, M., & Ambroise, C. (2004). Discriminative classification *vs* modeling methods in CBIR. *IEEE Advanced Concepts for Intelligent Vision Systems (ACIVS)*. Brussel, Belgium.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The element of statistical learning*. Springer.

Joachims, T. (1999). Transductive inference for text classification using support vector machines. *Proc. 16th International Conference on Machine Learning* (pp. 200–209). Morgan Kaufmann, San Francisco, CA.

Lewis, D., & Catlett, J. (1994). Heterogenous uncertainty sampling for supervised learning. *International Conference on Machine Learning* (pp. 148–56).

Lindenbaum, M., Markovitch, S., & Rusakov, D. (2004). Selective sampling for nearest neighbor classifiers. *Machine Learning*, 54(2):125–152.

Najjar, N., Cocquerez, J., & Ambroise, C. (2003). Feature selection for semi supervised learning applied to image retrieval. *IEEE ICIP*. Barcelena, Spain.

Park, J. (2000). On-line learning by active sampling using orthogonal decision support vectors. *IEEE Neural Networks for Signal Processing*.

Roy, N., & McCallum, A. (2001). Toward optimal active learning through sampling estimation of error reduction. *International Confenrence on Machine Learning*.

Smola, A., & Scholkopf, B. (2002). *Learning with kernels*. MIT Press, Cambridge, MA.

Tong, S., & Chang, E. (2001). Support vector machine active learning for image retrieval. *ACM Multimedia*.

Tong, S., & Koller, D. (2000). Support vector machine active learning with applications to text classification. *International Conference on Machine Learning*, 999–1006.

Tong, S., & Koller, D. (2001). Support vector machine active learning with applications to classification. *The Journal of Machine Learning Research*, 2:46–66.

Vasconcelos, N. (2000). *Bayesian models for visual information retrieval*. Doctoral dissertation, Massachusetts Institute of Technology.

Vasconcelos, N., & Kunt, M. (2001). Content-based retrieval from image databases: current solutions and future directions. *International Conference in Image Processing* (pp. 6–9). Thessaloniki, Greece.

Veltkamp, R. (2002). *Content-based image retrieval system: A survey* (Technical Report). University of Utrecht.

Zhu, X., Ghahramani, Z., & Lafferty, J. (2003). Semi-supervised learning using gaussian fields and harmonic functions. *International Conference on Machine Learning*.

Ant-like mobile agents for Content-Based Image Retrieval in distributed databases

Arnaud Revel REVEL@ENSEA.FR
David Picard DAVID.PICARD@ENSEA.FR
Matthieu Cord CORD@ENSEA.FR
Equipe Traitement des Images et du Signal
ENSEA/UCP/CNRS UMR 8051
6, avenue du Ponceau,
95014 Cergy Cedex France

Abstract

In this demo we present a Multi-Agent System using mobile agents for Content-Based Image Retrieval on distributed databases. The search system is based on 2 sub-systems: the databases discovery, using "ant-like" agents marking their route; the database indexing and retrieval using color signatures of the images.

Content-Based Image Retrieval (CBIR) are intrinsically limited by the semantic gap: the low-level information extracted from images and the semantic user request are very different by nature (Santini et al., 2001; Eakins, 2002; Smeulders et al., 2000). The final user is more interested in the semantic content than by the color or the texture of the image... although this is what is actually used by CBIR techniques. Besides, databases are now distributed all over the Internet.

One strategy to reduce this gap is to allow on-line interactive search (Ishikawa et al., 1998; Tong & Chang, 2001; Cord et al., 2004; Gosselin & Cord, 2004). The system asks the user to conduct the search within the database. Starting with a coarse query, the interactive process allows the user to refine the query as much as necessary. Most of the times, the user interaction consists of binary annotations indicating whether the image is relevant or not. The system integrates these annotations through a relevance feedback step to improve the system effectiveness.

Currently, interactive strategies only address fixed, closed image databases. Besides, the database is centralized. In the demo, we extend the problem to distributed databases (image databases distributed on a local network) by using Multi-agent systems (MAS) whose interest has been shown concerning resource sparing (Bonabeau et al., 2000; Simonin et al., 1998).

In previous works (Revel, 2003) inspired from robotics (Revel et al., 1998; M.Quoy et al., 2000; A.Revel & P.Gaussier, 2003), a MAS based on ANT algorithms (M.Dorigo et al., 1996) was proposed to find text content in HTML pages on the web. In this demo, we show how the system has been adapted to find images.

1. Summary on the "ant" search web-agent

The principle of "ant" strategies is to optimize routes towards a given resource by reinforcing markers (called pheromones) on sites situated along the pathway from the source to the destination. Given a set of agents launched from site s_1 (source) and which should reach site s_n (destination), the optimization is performed collectively and is the result of the emergence of a dynamical attractor coming from the interaction between all the agents and their environment.

Formalization:

Let $Ph_k(t)$ be the pheromone level on site s_k.

Let $\{s_1, ..., s_k, ...s_n\}$ be the set of sites

Let $prof$ the number of moves since s_1

Let $\{succ(s_i)\}$ be the set of sites s_j directly following s_i

Let $\{pred(s_i)\}$ be the set of sites s_j the agent has visited before reaching s_i

Let ξ be some noise value $\in [0, 0.0001]$

Practically, we have proposed the following algorithm:

Appearing in *Proceedings of the workshop Machine Learning Techniques for Processing Multimedia Content*, Bonn, Germany, 2005.

"Ant"-agent A behavior

Do
  If A is on site s_n
    `// The information is found!`
    Increase Ph_n
    Go back to the nest & Increase Ph_j $(\forall j \in \{pred(s_n)\})$
  Else
    `// Look forward`
    Decrease Ph_j $(\forall j \in \{succ(j)\})$
    Go to site s_j with probability $\frac{Ph_j(t)}{\sum_{l=1}^{k} Ph_l(t)} + \xi$
  End If
While s_n not found or $prof > \theta$

With Decrease and Increase given by:

Pheromone Ph_k updating

Decrease: $Ph_k(t+1) = (1-\alpha) \cdot Ph_k(t)$

Increase: $Ph_k(t+1) = Ph_k(t) + \beta$

Given this algorithm, we have shown that a population of "ant-agents" is able of optimizing a route leading to a given information and exhibit re-routing abilities.

2. Setup of the Demo

Figure 1. top) Image used as request — bottom) Example of answers given by the system.

In the demo, the agents are only able to move to three different machines directly from the source of the research (user interface — see 2). The first (A), is filled with 100 *animals* images, the second with 100 *flowers* images and the third contains both the animals and the flowers images (200 images[1]). Each research uses 20 "ant"-agents simultaneously (see figure 3). We choose these types of images because of their "orthogonal" characteristics: animals are very different from flowers both in meaning and in color composition. A signature for the target image (see figure 1) is computed and is compared to the signatures of the images within the database. When the distance is under a given threshold, the corresponding images is considered as correct (smiling "smileys" vs sad "smileys"). As computing the signatures could take much time and thus cannot be performed "on-the-fly", we have chosen to compute an index of the signatures of all the images "off-line" and stored it in a file containing both the URL of the images and the corresponding signature. The features we used to compare the example to the images from the databases were based on a HSV transform. The HSV space is known for its similarities with humans color perception. The HSV representation of the image was quantified in 163 colors, and the histogram was compute as a signature (Smith & Chang, 1996).

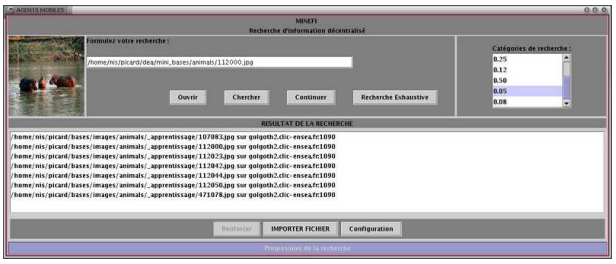


Figure 2. User interface: the user can select the image to search (upper left corner), the threshold (upper right corner). The URL of the results are given on the bottom of the window.

Two images are considered as alike when the euclidean distance between their signatures is under a threshold fixed by the user (see figure 2 and example of results in figure 1).

$$\sum_{1 \leqslant i \leqslant 163} |c_i(r) - c_i(k)|^2 \leqslant s \qquad (1)$$

The user interface asked the user to choose an image example and compute its HSV histogram. It also asked

[1] Although the demo uses 200 images only for purpose of simplicity and resources sparing, we are currently using databases with 1200 images concerning 14 categories.

for the threshold and save it with the histogram in a string. The string was then sent to the "ant"-agents.

The agent making the interface with the image database receives the string containing the signature and the threshold. It parses an indexes the file containing the histograms of the images stored on the computer and computed the similarity between the received signature and those in the index file (see equation 1). The URL of the images of which the histogram comparison were under the threshold are finally returned to the "ant"-agent coming back to the user (see figure 2).

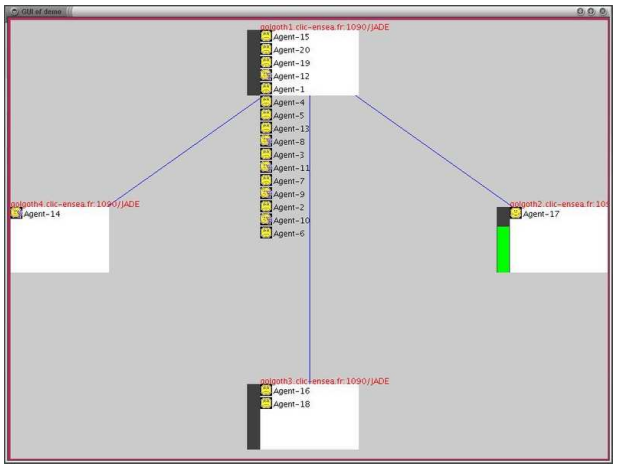


Figure 3. Demo interface: agents are launched from the upper platform and can reach 3 different machines (left, bottom and right). If an image is found (right "smiling" smiley), the agent comes back and reinforces the pheromone level (grey/green bar). Conversely, if no image is found (bottom "sad" smileys), the pheromone level decreases (dark bar).

References

A.Revel, & P.Gaussier (2003). *Biologically inspired robot behavior engineering*, chapter Designing neural control architectures for an autonomous robot using vision to solve complex learning tasks, 299–350. Springer-Verlag.

Bonabeau, E., Dorigo, M., & Theraulaz, G. (2000). The social insect paradigm for optimization and control. *Nature*, *406*, 39–42.

Cord, M., Fournier, J., Gosselin, P., & Philipp-Foliguet, S. (2004). Interactive exploration for image retrieval. *EURASIP Journal on Applied Signal Processing (special issue: Advances in Intelligent Vision Systems: Methods and Applications), accepted for publication.*

Eakins, J. (2002). Towards intelligent image retrieval. *Pattern Recognition*, *35*, 3–14.

Gosselin, P., & Cord, M. (2004). RETIN AL: An Active Learning Strategy for Image Category Retrieval. *IEEE International Conference on Image Processing (ICIP).* Singapore.

Ishikawa, Y., Subramanya, R., & Faloutsos, C. (1998). MindReader: Querying databases through multiple examples. *Proc. 24th Int. Conf. Very Large Data Bases, VLDB* (pp. 218–227).

M.Dorigo, Maniezzo, V., & Colorni, A. (1996). The ant system: Optimization by a colony of cooperating agents. *IEEE Transactions on Systems, Man, and Cybernetics-Part B*, *1*, 29–41.

M.Quoy, Gaussier, P., Leprtre, S., Revel, A., Joulain, C., & Banquet, J. (2000). *Lecture notes in artificial intelligence*, vol. 1812, chapter A planning map for mobile robots: speed control and paths finding in a changing environment, 103–119. Springer.

Revel, A. (2003). Web-agents inspired by ethology: a population of "ant"-like agents to help finding user-oriented information. *IEEE WIC'2003 : International Conference on Web Intelligence.* (pp. 482–485). Halifax, Canada: IEEE Computer Society.

Revel, A., Gaussier, P., Lepretre, S., & Banquet, J. (1998). Planification versus sensory-motor conditioning: what are the issues ? *SAB'98* (pp. 129–138).

Santini, S., Gupta, A., & Jain, R. (2001). Emergent semantics through interaction in image databases. *IEEE Transactions on Knowledge and Data Engineering*, *13*, 337–351.

Simonin, O., Ferber, J., & Decugis, V. (1998). Performances analysis in collective systems. *ICMAS'98: 3rd International Conference on Multi-Agent Systems* (pp. 469–470).

Smeulders, A., Worring, M., Santini, S., Gupta, A., & Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*, 1349–1380.

Smith, J. R., & Chang, S.-F. (1996). Tools and techniques for color image retrieval. *Storage and Retrieval for Image and Video Databases (SPIE)* (pp. 426–437).

Tong, S., & Chang, E. (2001). Support vector machine active learning for image retrieval. *ACM Multimedia.*

Learning Human Motion Patterns from Symmetries

László Havasi HAVASI@DIGITUS.ITK.PPKE.HU
Péter Pázmány Catholic University, Piarista köz 1., H-1052 Budapest, Hungary

Zoltán Szlávik SZLAVIK@SZTAKI.HU
Analogic and Neural Computing Laboratory, Hungarian Academy of Sciences, P.O. Box 63, H-1518 Budapest, Hungary

Csaba Benedek BCSABA@SZTAKI.HU
Péter Pázmány Catholic University, Piarista köz 1., H-1052 Budapest, Hungary

Tamás Szirányi SZIRANYI@SZTAKI.HU
Analogic and Neural Computing Laboratory, Hungarian Academy of Sciences, P.O. Box 63, H-1518 Budapest, Hungary

Abstract

This paper outlines and demonstrates a new algorithm which is capable of extracting and classifying human walking from video image-sequences, even in the non-ideal case of typical varying illumination conditions. The method works with spatio-temporal input information to detect and classify the patterns typical of human movement. The paper presents a new information-extraction and temporal-tracking method based on a simplified version of the symmetry which is characteristic for the moving legs of a walking person. In a further processing stage these patterns are filtered, then re-sampled using Bezier-splines to generate an invariant and noise-cleaned signature or "feature". With this use of temporal tracking and non-linear classification we have achieved pedestrian detection from cluttered image scenes with a correct classification rate of 91% from 1-2 step periods. The detection rates of linear classifier, SVM and Gaussian mixture model are also presented in the eigenspace for recognition purposes.

1. Introduction

Gait analysis has a wide range of possible applications. Being a no intrusive biometric feature it can be used for recognition/detection of people. Gait can be detected at low resolution, and therefore it can be used in situations where face or iris is not available in high enough resolution for recognition. The impact of pedestrian detection in intelligent vehicles could be great. The extraction and the tracking of human figures in image-sequences is a key issue for video surveillance and video-indexing applications. This interest is strongly motivated by the need for automated person-identification systems (http://www.darpa.mil/iao/HID.htm). The process may be broken down into several parts: detection (Cutler & Ellis, 2000), tracking, classification (Murase & Sakai, 1996) and identification (Wang et al., 2003, Hayfron-Acquah et al., 2003) of human movement or gait.

Two main trends are pursued in recent research on gait analysis. One takes into account motion information and tries to detect the periodic features of human gait in the movement of candidate objects. The second trend does not integrate information from previous frames, but rather verifies possible pedestrians by means of shape/silhouette analysis or pattern matching. The more traditional approach to the classification of a periodic signal is with the Fourier transform. Some systems operate a frequency analysis of the candidate pattern changes over time and then select the one characteristic of human gait. The selection is made with statistical methods or by simple thresholding (Fujiyoshi&Lipton 1998, Lipton 1999, Polana&Nelson 1997). As an example in (BenAbdelkader et al. 2001, Cutler&Ellis, 2000) a Short-Time Fourier Transform is used with a Hanning windowing function to analyse the signals obtained by correlation of the pattern of detected objects. HMM based models are very popular for human gait analysis. In (Kale et al.) the periodic change of human's silhouette was modeled by a HMM. In (Curio et al., 2000) the periodic movement detected is correlated to an experimental curve derived from the statistical average of human gait periods. High peaks of the correlation function indicate the presence of a person. A useful and popular approach is based on silhouette

Appearing in *Proceedings of the workshop Machine Learning Techniques for Processing Multimedia Content*, Bonn, Germany, 2005.

analysis (Soriano et al. 2004) with spatiotemporal representation, where a key aim is to achieve a more invariant data representation of the detected object. In the work of Hayfron-Acquah et al. (2003) the symmetries of the silhouette are utilized as a biometric parameter for person-identification.
Our method focuses on the periodic dynamics of human gait. It utilizes derived third-level symmetries of the edge-map to detect and track structural changes in video sequences, and uses SVM for pattern classification in an eigenspace with a small number of dimensions (it uses PCA to reduce the dimensionality of the feature-space). Our method (under certain not severely restrictive assumptions) can detect pedestrians in real-time, with 91% correct detection.

2. Feature Extraction

Generally, object detection in videos must be done in two steps. Firstly, some detectable features are needed; secondly, robust tracking of the extracted features must be performed.
The present paper describes a method, comprising the following processing steps on the input image:

- Adaptive background filtering
- Constructing third-level symmetries
- Tracking the symmetries
- Re-sampling motion patterns
- Dimension reduction of normalized patterns
- Pattern classification: walk or non-walk

2.1 Background Modeling

For the detection of changes in video image sequences we have implemented an adaptive background-modelling algorithm. Our method tries to extract the accurate silhouettes of foreground objects even if they have partly background like colors and shadows are observable on the image. It does not need any a priory information about the shapes of the objects, it assumes only they are not point-wise. The method exploits temporal statistics to characterize the background and shadow, and spatial statistics to the foreground. A Markov Random Field model is used to enhance the accuracy of the separation.
The background modelling step is based on the work of (Stauffer et al. 2000). The algorithm collects statistics about the occurring values at each pixel position, and the recent history is modeled as a mixture of Gaussians. The i-th component of the mixture has the following parameters: weight w^i, mean value μ^i and covariance matrix Σ^i in $\Sigma^i=\sigma^i \cdot I$ form. (I is the 3x3 identity matrix.) The component with the greatest weight is considered the background component. The parameters are updated via online k-means algorithm.
Shadow detection is an important issue. Usually shadows have to be handled separately, because they do not belong to moving objects but their color properties are different from the background, see Figure 2. Our shadow detector works in the HSV color space. (Cucchiara et al. 2001) established that shadow cast on the background does not change its hue significantly, but it lowers its luminance. The detector uses the parameters of the current background model. The output I is a binary decision:

$$I(s)=\begin{cases} 1, \ if \ \left|x_s^H-\mu_0^H(s)\right|<\tau_1 \ \ \& \ \left|x_s^V-(\mu_0^V(s)-\tau_{offset})\right|<\tau_2 \\ 0, \quad otherwise \end{cases}$$

Parameters $\tau_{offset}, \tau_1, \tau_2$ are characteristic for lighting conditions and reflection.
For the detection we used a general Markov Random Field model described in (Berthod et al. 1996). Each pixel is classified into one of the following classes: foreground, background, and shadow; and a label is assigned to it according to the classification result. An energy term constructed depending on the global labelling of the entire image. One part of this energy term evaluates the similarity between the current value of the pixel and the "usual value" in the actual class of the pixel. The other term penalizes the high number of neighbouring pixels having different label. The aim is to minimize the energy term, so the algorithm should find the best, or nearly the best labelling in the image.
We developed a probability model to describe the desired classes and get the first part in the energy term. We collected the background, and shadow statistics in time using the previously mentioned methods, and the foreground properties in space through a pixel value classification algorithm. We introduced the details of our work in (Benedek&Sziranyi 2005).

Figure 1: a) Original frame b) Silhouette image created by Stauffer-Grimson algorithm c) Foreground and shadow detection with our method (white: foreground, gray: shadow, black: background)

Figure 1 shows the result of our method in a difficult case when a man is walking in white shirt in front of a white wall.

Figure 2: Experimental results: a) a good detection results b) detection problem caused by reflection and shadows

2.2 Local-Symmetry Extraction

Symmetry is a basic geometric attribute, and most objects have a characteristic symmetry-map (Giblin & Kimia, 2003). These unique and invariant properties lead to the applicability of symmetries in our approach to image-processing. Our method (Havasi & Szlávik, 2004) employs a modified shock-based method (Sharvit et al., 1988): it calculates symmetries by propagating parallel waves from the ridge. In our approach, we simplify the algorithm by using only horizontal morphological operators; since, in the practical cases we are considering, we essentially need to extract only vertical symmetries. This modification has the advantage that it assists in reducing the sensitivity to fragmentation. Sample outputs of the algorithm can be seen in Figure 4. The symmetry operator normally uses the edge map of the original image as its input; we used the Canny edge-detector algorithm to derive the locations of the edges (ridges).

Figure 3: An idealised outline of a walking person, together with the derived Level 1, Level 2, and Level 3 symmetry maps.

As illustrated in Figure 3, the symmetry concept can be extended by iterative operations. The symmetry of the Level 1 symmetry map is the Level 2 symmetry; and the symmetry of the Level 2 map is the Level 3 symmetry (L3S).

2.3 Tracking

The extracted L3Ss are primary useful for analysing images of the legs of the human target. On the other side, symmetry-images resulted from the arms are typically composed of small fragments which are difficult to distinguish from the noise. However, even the existence of clear symmetries in a single static image does not necessarily provide usable information about the image content; for this, we need to track the changes of the symmetry fragments by temporal comparisons. The symmetry fragments and their radii define an outline that can be used as a mask between successive frames to aid classification of the coherent fragments in successive frames, as illustrated in Figure 4. The tracking algorithm calculates the overlapping areas between symmetry masks; and as time progresses it constructs the largest overlapping area. The results of temporal tracking can be seen in Figure 4, where we demonstrate the resulting symmetry-patterns in real-life situations. The first L3S appears when the legs are opening and the last is detected just before the legs are closed; so a symmetry-pattern for the motion of a walking person corresponds to the movement of the legs from opening to closing.

Figure 4: Masks of the reconstructed symmetries from successive frames; and symmetry-patterns.

2.4 Re-Sampling

The extracted symmetry-patterns are represented with the upper and lower end points (2 each) of the L2S in each frame; these points can be seen on Figure 5.

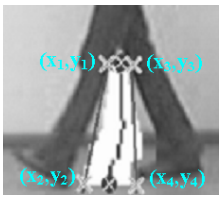


Figure 5: The limits used to define symmetries for the re-sampling and classification tasks.

Thus there are four 3D coordinates, which correspond approximately to the "end-points" of the two legs. Temporally these patterns depend both on the frame rate and the walking speed, so a pattern usually contains data from 4-8 frames. Rates of digital cameras may vary by time when bus-control drops the transmission. It is not rare and it needs the re-sampling on time. We perform this normalization task with Bezier spline interpolation. This technique has the advantage that it performs two tasks: (i) data is re-sampled in a defined time interval with fixed point count, (ii) noise-filtering is performed on the trajectories, which gives a smoother symmetry-pattern. In our experiments the noise-cleaning is critical because in real outdoor scenes these patterns are often damaged. The Bezier spline (B-B spline) (1) is a good choice because the effect of base points is global; so the presence of some damaged points does not cause significant change in the whole trajectory.

$$\vec{r}(t) = \sum_{i=0}^{n} B_{i,n}(t)\vec{P}_i$$
$$B_{i,n}(t) = \binom{n}{i} t^i (1-t)^{n-i} \quad (\mathrm{i} = 0,1,\ldots,\mathrm{n}) \tag{1}$$

The base points consist of the (x,y) image coordinates of the above-mentioned symmetry patterns, while the z coordinate represents the time (in milliseconds) at which the frame was captured. This time-extended data representation permits the integrated analysis of data

obtained from several cameras where the frame rate is different (e.g. network cameras); the extracted features must be resampled with a continuous time-division. The result of Bezier spline interpolation of data can be seen in Figure 6.

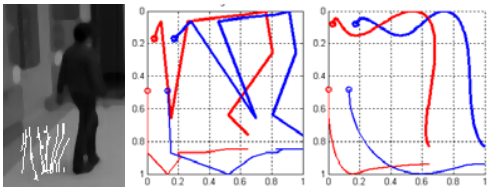


Figure 6: Original symmetry-pattern, the trajectories of 9 frames and interpolated trajectories of 100 points.

2.5 Dimension Reduction

The interpolated 3D (XYT) points are rearranged into a row-vector with dimension of 800:

$$\tilde{x} = [x_1, y_1, x_2, y_2, x_3, y_3, x_4, y_4]$$

It follows from the linearity of the z coordinate (time), the smooth time-division (time is linearly related to the successive samples), that this coordinate can be omitted: it has no discriminative information-content. After the patterns are centered in both x and y, both coordinates are normalized using a constant chosen such that max(y)= 1 and min(y)= –1; we do this because we have found that the y-size of the patterns varies less than does the x-size. We do not normalize with individual coefficients for x and y, since in that case the information-content of the ratio of the x and y values would be lost.

A well-known technique for dimension reduction is the PCA method (Huang et al., 2001). To find the principal components of the distribution of the feature space we first obtain the mean m and the covariance matrix Σ of the data set. Then we can compute $N \leq rank(\Sigma)$ nonzero eigenvalues and the associated eigenvectors of Σ based on SVD. The eigenvectors associated with a small number of the largest eigenvalues correspond to large changes in training patterns; thus a transformed matrix can be constructed from eigenvectors to project the original data into a parametric eigenspace with a drastically reduced number of dimensions. We keep the first 3 eigenvalues and their associated eigenvectors to form the eigenspace transformation matrix. Figure 7 gives these three eigenvectors. From Figure 7, we can see that these eigen-walks are periodic, which reveals the construction method of raw data. Furthermore we can determine that the relevant information is the x directional motion of lower end points and walk has a characteristic symmetry on the y directional motion of end points.

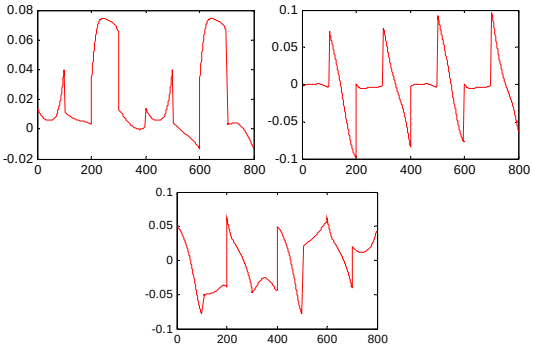


Figure 7. The first three eigenvectors obtained by PCA training.

We considered the space spanned by the 3 most significant eigenvectors that account for 93% of the variation in the training data-set: we call this the Eigenwalk space. Figure 8 demonstrates the results using the test-set of labeled "walk" and "non-walk" patterns which is described below.

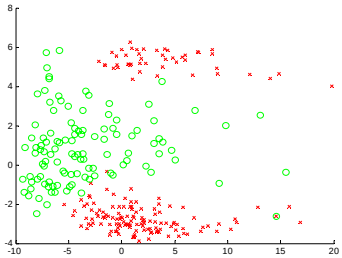


Figure 8. "Walk" and "non-walk" patterns in the eigenspace

In the Figure 8 we can observe two main groupings of the "walk" patterns; these two groups correspond to the two opposite walking-directions. The human patterns lie on a non-linearly shaped manifold in the eigenspace. This drastically reduced number of dimensions is of great assistance in increasing the speed of classification, which is an important factor in real-time applications.

3. Pattern Classification

Level 3 symmetries can also appear in other parts of the image, not only between the legs; and the tracking method also collects all of these related symmetries. Our previous work (Havasi et al., 2004) introduced an algorithm that is able to detect pedestrians from recognition of their characteristic symmetry-patterns, using Kernel Fisher Discriminant Analysis (KFDA technique). Here we present a more established pattern-classification method based on the continuous interpolation of the symmetry-patterns and the classification process is carried out via non-linear classification method, namely Support Vector Machine (SVM) (Müller et al., 2001).

In our experiments we compared the performance of linear – Linear Discriminant Analysis (LDA), Linear Perceptron, and non-linear – KFDA with gaussian kernel,

SVM with gaussian (2) and inverse multiquadratic (3) kernels, classification methods. Selecting properly the kernel parameters for SVM the number of support vectors can be controlled with an acceptable classification error rate, see Figures 9 and 10. The number of support vectors in the SVM-based algorithm has a direct effect on the speed of the algorithm: less the number of support vectors faster is the algorithm.

$$k_\sigma(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) \quad (2)$$

$$k_\theta(x,y) = \frac{1}{\sqrt{\|x-y\|^2+\theta^2}} \quad (3)$$

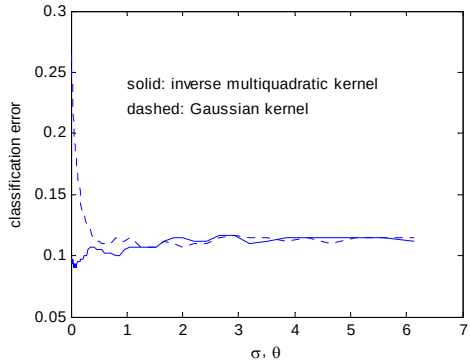


Figure 9. Relation between the kernel parameter and the classification error rate for the Gaussian (2) and inverse multiquadratic (3) kernels

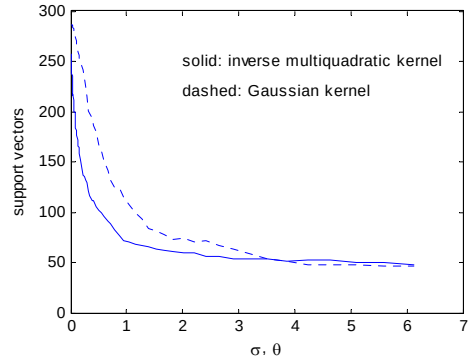


Figure 10. Relation between the kernel parameter and the number of support vectors

Analyzing Figures 9 and 10 we obtained that the optimal kernel parameters are σ=1.26 and θ=0.92 for gaussian and inverse multiquadratic kernels respectively.

To evaluate the proposed method, we derived "walk" and "non-walk" patterns from a considerable number of real-life indoor and outdoor video sequences representing a variety of different walk directions, frame rates, viewing distances and surrounding situations. For training we used 920 samples, and according to our manual classification these comprised 420 "walk" and 500 "non-walk" patterns. In the experiments our main goal was to reliably detect human movements, but at the same time with a false-positive detection rate as small as possible. Testing of the methods were made on a dataset of 1500 samples, 500 "walk" and 1000 "non-walk" patterns. Table 1 summarizes the detection results of the tested classification methods.

Table 1. Correct classification rates of different methods

METHOD	DETECTION RATE	FALSE-POSITIVE	FALSE-NEGATIVE
LDA	76.4%	15.6%	8.0%
KFDA (GAUSSIAN)	87.9%	7.2%	4.9%
LINEAR PERCEPTRON	79.0%	17.3%	7.7%
SVM (GAUSSIAN)	89.2%	8.2%	2.6%
SVM (INV. MULTIQUAD.)	91.0%	6.9%	2.1%

4. Experimental Results

After considering the numerical results, we summarize some practical limitations (see Figure 11) of the symmetry-tracking method which we noted. The L3Ss can be evolved only if the leg-opening is visible. In our tests we found that this meant that the direction of movement had to be at more than about 70° from the viewing axis; but this is not a serious limitation when more than one camera is monitoring the area (Szlávik et al., 2004). Crowds, and some other specific "overlap" situations are the main cases which cause problems, although "overlap" does not always prevent successful tracking. The most common problematic cases were as follows: subject wearing long coat; subject carrying large bag etc. in the hand nearest to the camera; full masking of the legs by another person in the perspective view; partial masking by another person moving on a parallel track, with synchronized step periods. All in all, the proportion of such non-conform walking (with no visible leg-pairs) in the processed real-life video sequences was about 20% in the campus area.

Figure 11. Typical problematic cases

5. Conclusions

The method we describe can detect pedestrians in image-sequences obtained in outdoor conditions in real-time.

Considering even a single step-period, a very low false-detection rate is obtainable. To achieve this, we used a novel feature-extraction and tracking method that can reflect the natural structural changes of human leg-shape; the method seems promising for the purpose of providing a useful "understanding" of image-content and thanks for the simplified symmetry extraction algorithm the detection method can run 15-20 FPS in 640x480 resolution on a 2.4GHz Pentium CPU.

In this paper, we have introduced an extended version of our pedestrian detection method described in a previous paper (Havasi et al., 2004). We are able to achieve an improved detection rate because we now use a more invariant and effective data representation in the Eigenwalk space, based on spline interpolation and a dimension-reduction technique.

The method appears suitable for the detection of human activity in images captured by video surveillance systems such as those typically used in public places.

Acknowledgement

The authors would like to acknowledge the support received from the NoE MUSCLE project of the EU.

References

Cutler R., and Ellis T. (2000), Robust real-time periodic motion detection, analysis and applications. *IEEE Trans. on PAMI,* vol. 22, pp. 781-796.

Murase H., and Sakai R. (1996), Moving object recognition in eigenspace representation: gait analysis and lip reading *Pattern Recognition Letters,* vol. 17.

Wang L., Tan T., Ning H., and Hu W. (2003), "Silhouette Analysis-Based Gait Recognition for Human Identification" *IEEE Trans. PAMI,* vol. 25, pp. 1505-1518.

Soriano M., Araullo A., Saloma C. (2004), "Curve spreads-a biometric from front-view gait video" *Pattern Recognition Letters,* vol. 25, pp. 1595-1602.

Hayfron-Acquah J., Nixon M., and Carter J. (2003), "Automatic gait recognition by symmetry analysis" *Pattern Recognition Letters,* vol. 24, pp. 2175-2183.

Curio C., Edelbrunner J., Kalinke T., Tzomakas C., W. von Seelen (2000), "Walking pedestrian recognition" *IEEE Trans. Int. Transport Systems,* pp. 155-163.

BenAbdelkader C., Cutler R., Nanda H., and Davis L. (2001), "Eigengait: Motion-based Recognition of People Using Image Self-Similarity" *Proc. Int. Conf. Audio- and Video-Based Biometric Person Authentication*, pp. 284-294.

Stauffer C., Eric W., and Grimson L. (2000), Learning patterns of activity using real-time tracking, *IEEE Trans. on PAMI, 22(8),* 747-757.

Havasi L., Szlávik Z. (2004), Symmetry feature extraction and understanding, *Proc. CNNA'04*, pp. 255-260

Sharvit D., Chan J., Tek H. and Kimia B.B. (1988), Symmetry-based indexing of image databases, *J.Vis. Comm. and Image Represent.*, vol. 9 no. 4, pp.366-380

Huang P., Harris C., and Nixon M. (2001), "Human Gait Recognition in Canonical Space Using Temporal Templates" *IEE Proc. Vision Image and Signal Processing Conf.*

Havasi L., Szlávik Z., and Szirányi T. (2004), „Pedestrian detection using derived third-order symmetry of legs", *ICCVG*.

Müller K.-L., Mika S., Ratsch G., Tsuda K., and Schölkopf B. (2001), "An Introduction to Kernel-Based Learning Algorithms," *IEEE Trans. On Neural Network*, 12(2).

Szlávik Z., Havasi L.,and Szirányi T. (2004), Estimation of common groundplane based on co-motion statistics, *ICIAR, Lecture Notes on Computer Science*

Fujiyoshi H., Lipton A. (1998), Real-time human motion analysis by image skeletonisation, *IEEE Proc. WACV.*

Lipton A. J. (1999), Local application of optic flow to analyse rigid versus non rigid motion, *ICCV*.

Polana R., Nelson R. C. (1997), Detection and recognition of periodic, non-rigid motion, *Int.J.Comp. Vision*, 23(3).

Kale A., Rajagopalan A.N, Sunderesan A., Cuntoor N., Roy-Chowdhury A., Krueger V., Chellappa R. (2004) Identification of Humans Using Gait, *IEEE Trans. on Image Processing*, Sept . 1163-1173.

Giblin P. J., Kimia B. B. (2003), On the Intrinsic Reconstruction of Shape from Its Symmetries, *IEEE Trans. PAMI,* vol. 25, pp. 895-911.

Cucchiara R., Grana C., Neri G., Piccardi M., and Prati A., (2001) The Sakbot System for Moving Object Detection and Tracking, *Video-Based Surveillance Systems—Computer Vision and Distributed Processing*, pp. 145-157.

Berthod M., Kato Z., Yu S., Zerubia J. (1996): "Bayesian image classification using Markov Random Fields". *Image and Vision Computing* 14: 285-295.

Benedek Cs., Sziranyi T. (2005) "Markovian framework for foreground-background-shadow separation of real-world video scenes" *EMMCVPR 2005*, under review

Addressing Partial Relevance in Image Retrieval through Aspect-Based Relevance Learning

Mark Huiskes
Centre for Mathematics and Computer Science (CWI), Amsterdam

Abstract

We consider the special structure of the relevance feedback learning problem in image retrieval, focusing particularly on the effects of image selection by partial relevance on the clustering behavior of feedback examples. Aspect-based relevance learning addresses this issue directly by means of a hypothesis testing approach. We evaluate its performance by comparison to two feature re-weighting methods.

1. Introduction

As image content interpretation is both user- and task-dependent, content-based image retrieval (CBIR) revolves to an important extent around the task of *interactively* reaching an understanding of what a user is looking for. One natural type of interaction is by soliciting feedback directly in terms of the presented images: by analyzing indicated relevant (positive) and irrelevant (negative) example images, the system may iteratively improve the selection presented to the user. Feedback in terms of images is particularly convenient given that, unlike for text documents, relevance of images can truly be determined "at-a-glance". Recent reviews of the state-of-the-art of relevance feedback in CBIR are given in Zhang et al. (2003) and Zhou and Huang (2003).

As the importance of image features representing the image content differs from query to query, much research has been aimed at *feature re-weighting* (e.g. Rui et al. (1998), Salton and Buckley (1990)). For example, Rui et al. (1998) update weights of different feature classes by using the inverse variance of the positive examples, thereby giving higher weights to features for which the positives are relatively close together. Many variants of this approach have been proposed (e.g. Ciocca and Schettini (1999), Peng et al. (1999)) typically based on the idea of assigning higher weights to features in which positives cluster, while negatives remain separated.

In many recent approaches the feedback images are taken as training samples and are used to train a classifier or other learner for predicting the (ir)relevance of the database images. Examples of learning methods used are: SVMs (Tong & Chang, 2001), boosting (Tieu & Viola, 2004), decision trees (MacArthur et al., 2000), and nearest neighbors (Wu & Manjunath, 2001).

In this paper we also treat relevance feedback analysis primarily as a learning problem but as one with a special structure that requires careful attention. In Huiskes (2005a) we proposed aspect-based relevance learning as an analysis method well suited to this structure. In this paper the main contribution lies in the comparison of the performance of this method to two feature re-weighting methods, where we focus in particular on the effects of example selection by *partial relevance*. The experimental results are based on testing with a retrieval system for decoration designs (e.g. wallpaper or textile patterns) for which example selection based on partial relevance is a particularly pressing issue.

2. Structure of the Relevance Feedback Learning Problem

As a learning problem we cannot treat relevance feedback analysis as a standard two-class, relevant versus irrelevant, classification problem; we mention the following issues:

Small sample learning problem. It has often been recognized (e.g. Zhou and Huang (2003)) that the relevance feedback problem is a *small sample* learning problem. The number of example images depends on the willingness of the user to cooperate but is generally small, say at most 10 examples per feedback cycle, whereas the dimension of the feature space is large (often higher than 100). The small sample sizes disqualify many of the standard learning methods unless special measures are taken (e.g. Tong and Chang (2001)).

Example selection by partial relevance. Images are typ-

Appearing in *Proceedings of the workshop Machine Learning Techniques for Processing Multimedia Content*, Bonn, Germany, 2005.

ically relevant in some aspects and not relevant in others, and in many retrieval applications fully relevant images are hard to come by initially. When a user selects an image as feedback example he generally does so based on one or a few salient aspects; however, not all aspects of interest need to be present in the image, nor need all salient aspects present in the image be relevant. For features other than those by which an image is chosen, which can be a large majority, the feedback received is thus to a large extent random: positive feedback is given for feature values for which no such feedback was intended. As a consequence examples often provide *misleading evidence*, see Fig. 1.

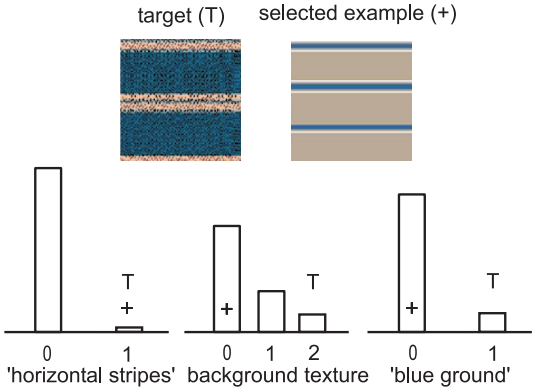


Figure 1. Shown are a target image (representing a simple user query for images of this type) and an image that the user has selected as a positive example. Also shown are histograms of database values for three (hypothetical) features: "presence of horizontal stripes", a feature measuring some characteristic of ground texture, and "presence of blue ground". The plus sign indicates the feature values of the example image; the T symbol indicates target values desired by the user. The example image is selected based on a single feature, viz. the possession of horizontal stripes. No positive feedback for other features was intended; as a consequence, such features will receive feedback on values that are (approximately) random draws from the feature value distributions. This often leads to misleading evidence, as is illustrated by the two other features shown here.

Examples will tend to cluster at feature values that are most common in the database, thus interfering with the identification of the proper regions of relevance. This is related to the next issue.

Feature value distributions. Features often have value distributions that are skewed. Take for instance a feature measuring the yellow-ness of an image, say divided into three classes: "no yellow", "some yellow" and "very yellow". Then most of the database images will be in the first class and, relatively, very few will be in the last. Generally only few images have values that correspond to perceptually salient properties.

The effect of example selection by partial relevance may thus be amplified by feature value distributions: clustering will naturally occur at the most common feature values. Even though negative examples may counteract misleading clustering of positives to some extent, learning methods will generally be influenced by the unintended concentration of positive examples and the relatively small fraction of feature values for which feedback was actually intended. This also holds for many feature re-weighting approaches as they are usually based on the variation or clustering behavior of example feature values.

3. Aspects and Relevance

In Huiskes (2005a) we treat images as sets of aspects, where we understand an "aspect" simply as a property which an image either has or has not, and for which we intend to resolve its effect on perceived relevance as a unit. Aspects can thus be explicitly defined in terms of conditions on feature values, i.e. as derived binary features that model a specific perceptual quality, but can also live solely in the "eye of the beholder".

There are two main reasons why we choose to employ aspects as an intermediary conceptual layer between the features and relevance estimates. First, it provides an effective framework for modeling partial relevance. Each aspect can be considered as either *neutral*, relevance enhancing (*positive*, or simply *relevant*) or relevance inhibiting (*negative*). In this way we can model a search task as a collection of positive, neutral and negative aspects. Note this is not the case for features as a whole. For "relevant features", not only will there be feature values that lead to higher perceived relevance, but by necessity there are also feature values making images less relevant. Second, it allows us to associate a frequency of occurrence to such "unit of relevance" given a specific context. As we understand the context to be the database under study, we define for each aspect an *aspect image frequency* p_{db} as the fraction of images in the database that possess the aspect. As explained in the next section this is the key to quantifying meaningful clustering and discerning neutral from positive and negative aspects. The actual construction of aspects is discussed in section 5.1.

For illustration, suppose a user is interested in finding designs that have: (i) a blue background; (ii) simple round motifs that are far apart; and (iii) high contrast between motifs and ground. Depending on the available features, we can translate this to requirements in terms of aspects. Some aspects are clearly positive,

e.g. blue-ness of the ground should be high, dominant motif shape should be round, and relative amount of background should be high. Aspects in opposition to relevant aspects are negative, e.g. the user does not want squares, or a ground that is red. Additional negative aspects may come up during the feedback process, e.g. the user may decide that he does not like yellow motifs. Other aspects are neutral, e.g. he may not care about ground pattern: it may be plain or have some texture.

4. Aspect-based Relevance Learning

In our retrieval system feedback example selection is implemented by presenting images in clickable selection displays, each consisting of a grid of a fixed number of, say 50, thumbnail images. The number of images inspected per cycle may be larger as the user can leaf through the selection displays, or "reset" for a new random selection. Additional selection displays may be available, for instance offering "most-informative-images" (e.g. Zhou and Huang (2003)). The sequential ordering of the images is random in the first cycle, and by relevance ranking in subsequent cycles. The examples and counterexamples are collected in positive and negative *example sets*. At each cycle of the feedback process the user updates the examples in the example sets by either: (i) selecting new images as positive or negative examples adding them to their respective sets; (ii) removing images from the sets, i.e. the sets are preserved unless images are no longer deemed representative enough and are deleted explicitly.

4.1. Aspect Selection

For aspect-based relevance learning we use the feedback data available at the end of each cycle foremost to establish the effect (neutral, positive or negative) of the various aspects. The main idea is the following: as the user selects an image as feedback example based on some positive or negative aspects, possession of the other aspects will approximately follow the distribution of aspect possession in the database. We are interested in finding those aspects for which the user has *actively* selected more examples with that aspect than may be expected to arise by chance only, i.e. as a side product of selection by other aspects. As for each aspect we know its associated aspect image frequency p_{db} we can model the probability distribution of the number of examples that would arise for a neutral aspect. Taking this approach has the benefit that feature selection and, ultimately, relevance assignment is based not only on clustering behavior of positives and negatives, but is also compared to clustering behavior of all database images. This leads to a natural emphasis on salient[1] aspects, effectively giving higher weights to example image feature values that are more rare in the database. In addition, by taking into account feature value distributions, we are not dependent on negative examples to down-weight positives that cluster at aspects with low saliency. This means negatives can be used to indicate which aspects are not desired, but are not required for the sole purpose of getting sufficient data for classification.

Let n^+ (n^-) be the total number of positive (negative) images selected, which we take to be fixed, and N^+ (N^-) be the number of positive (negative) examples that possess the aspect. For each aspect, we consider two *independence hypotheses*, H_0^+ and H_0^-, stating that the aspect behaves as if it were neutral to the user in regard to the accumulation of positive (resp. negative) examples. Under these hypotheses we model aspect possession of an example image as a Bernoulli variable with probability p_{db}; consequently, the number of positives and negatives with given aspect can be modeled as binomial variables with probability parameter p_{db}:

$$N^+ \sim B(n^+, p_{\mathrm{db}}), \quad \text{and} \quad N^- \sim B(n^-, p_{\mathrm{db}}). \tag{1}$$

We intend to select aspects as positive or negative, only if there is sufficiently strong evidence supporting this decision relative to the independence hypotheses. We do so by first assessing the probabilities of finding the same or a higher number of example images with the given aspect as in the current example sets given the aspect is neutral. If we select only those aspects for which these probability values are below a certain threshold, p_0^+ (resp. p_0^-), we limit the probability of the error of erroneously deciding that the aspect is not neutral.

More formally, we define two p-values associated with the respective hypotheses

$$p^+(N^+) = \sum_{i=N^+}^{n^+} \binom{n^+}{i} p_{\mathrm{db}}^i (1-p_{\mathrm{db}})^{(n^+-i)}, \tag{2}$$

with $p^-(N^-)$ defined analogously.

When we reduce the p-values, thereby raising the number of examples required for selection, we also increase the probability of missing actual positive and negative

[1] Saliency, in the sense of how rarely the aspect occurs in a given context, is inversely related to image frequency. Note that we do not use the tf/idf approach (e.g. Squire et al. (1999)): we do not have terms and use a rejection rather than a weighting mechanism.

aspects. As evidence is expected to accumulate in subsequent feedback cycles, we use the following dynamic p-value strategy. For the positive aspects we start with a relatively large p-value, say 0.05, in order not to miss relevant aspects when evidence is still relatively weak. After a number of feedback cycles (e.g. 3) evidence can be expected to have accumulated and the p-value is reduced, to say 0.001, in order to increase precision by avoiding false positive aspects. For negative aspects we take a small p-value (0.005) from the beginning, as negative feedback is necessary only when a certain aspect starts to accumulate in the display of highest ranking images, at which point sufficient examples will be available. To monitor evidence accumulation more accurately, explicit user involvement is required e.g. by indicating fully relevant examples or by measuring satisfaction with respect to the quality of the example sets.

4.2. Relevance Ranking

Let M be the aspect matrix with columns of boolean variables indicating if images have a given aspect or not. We can, trivially, determine N_j^+ and N_j^- from the image index sets S^+ and S^- of positive and negative examples, giving the two p-values, $p^+(N^+)$ and $p^-(N^-)$ by (2). Let A^+ be the index set of accepted enhancing aspects, and A^- be the index set of accepted inhibiting aspects, then the relevance rel_i for image i is defined by $\mathrm{rel}_i = \sum_j M(i, A_j^+) - \sum_j M(i, A_j^-)$.

Note that once a group of aspects is accepted, the decision of how take these into account of course need not be so black-or-white; a variety of weighting schemes could be devised to obtain more gradual aspect influences based on the strength of the evidence.

5. Query Simulation Experiments

At the end of each feedback cycle the information in the example sets is transformed into a new image relevance ranking. In the following we will compare the performance of relevance feedback inference mechanisms by means of query simulation experiments, i.e. by controlled simulation of example sets. As we explain below this allows us to simulate the "intention" behind an example set and to use this to compare the quality of the generated relevance rankings.

We simulate example sets by selecting a number of *target aspects*. Each aspect represents a feature value or range of feature values the user is actively interested in. The target aspects are randomly sampled from aspects with a p_{db} value below a given threshold. This assures the generation of example sets such that the target aspects are saliently present and perceptually relevant. In this study we consider only these positive aspects, and do not simulate negative aspects. The target aspects imply a set of *target images*, viz. those images that possess all the target aspects.

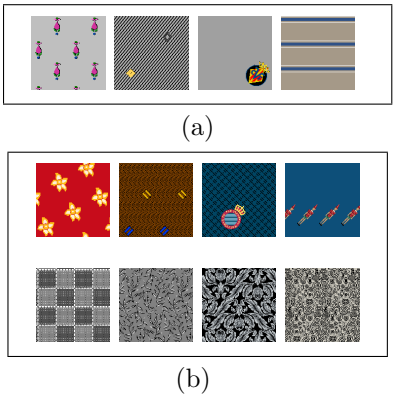


Figure 2. Two simulated example sets, for target aspects "large amount of background" and "very grey", generated according to the (a) full relevance scenario, and (b) partial relevance scenario.

We consider two main scenarios for generating the positive example sets. In the *full relevance* scenario the positive example set consists of a fixed number of randomly sampled target images. Fig. 2 (a) shows a generated example set of 4 target images for two target aspects, labeled as "large amount of background" and "very grey". In the *partial relevance* scenario we generate a fixed number of images for each of the target aspects, randomly sampled from those images that have such aspect but do not have the remaining target aspects. In Fig. 2 (b) an example set is shown that is generated for the same two aspects as before but now according to this scenario. Note for instance that the last 4 images are grey but have relatively little background. Compared to the first scenario where simulated example sets consist strictly of fully relevant images, this scenario represents the other extreme where feedback is provided exclusively by means of partially relevant images. The scenarios represent realistic modes of relevance feedback as may occur in the image retrieval process; for both cases users would expect rankings for which images with target aspects are ranked higher. As further detailed below we quantify performance by means of precision-recall graphs for the target images, as well as by counting the number of target aspects present in the top-ranking images.

We have compared the aspect-based relevance learning method to two feature re-weighting methods. These methods represent an interesting reference of performance as they are also based on clustering of examples, but do not take into account evidence if such

clustering is significant given the distribution of feature values in the database. The first re-weighting approach (FRW1) is that of Ciocca and Schettini (1999) in which feature weights are inversely proportional to the mean normalized distance within the example set. Higher weights are thus given to features for which the examples are similar. The second re-weighting approach (FRW2) follows Rui et al. (1998) where feature weights are taken inversely proportional to example variance. For both approaches, the relevance rankings are based on the moving query point mechanism (e.g. Rocchio (1971)). Each example set determines an "ideal query point", computed as the average feature vector over the examples, where in our case all examples receive equal weight. The relevance ranking then follows from the sorted distances of the images to this query point, using the weights of the features determined as described above. For the aspect-based relevance learning method we consider two different p-values, ARL1 using $p = 0.05$, and ARL2 using $p = 0.001$, respectively.

5.1. Aspect Construction for Decoration Designs

Testing is based on aspects and features for a database of decoration designs. To characterize decoration designs we have selected and developed a variety of features suitable for representing their global appearance; these include features for: color, texture, complexity and periodicity. In addition several features have been computed based on the decomposition of designs into figure and ground, e.g. relative amount of background, background texture, properties of motifs (e.g. size, number, variation) and their spatial organization. Finally, a set of 42 manually annotated semantic category labels (e.g."geometric") is also available. For details we refer to Huiskes (2005b).

Construction of aspects varies by feature type. Binary and discrete features can be converted directly into aspects. For single dimensional numerical features we use quantization, either manually by inspection, or automatically. We have taken an automatic approach based on a grouping mechanism: we take a redundant group of aspects, defined at a number scales and overlapping in range (here a total of 14 aspects at 2 scales), and consider for each scale only the aspect with the smallest p-value of the group as a candidate for selection. High aspect redundancy is feasible as the computational cost per aspect is very small. For higher dimensional feature spaces our preferred solution is to take an exemplar or case-based approach. For instance, we have selected a number of simple example shapes as prototype shapes, and defined a "simple-motif" aspect by marking shapes that are close enough to one of the prototypes based on the similarity metric of the MPEG-7 contour shape descriptor. Another approach would be to construct data-driven aspects by mining for image clusters in feature spaces, where aspects again follow from cluster membership.

Numerical features were computed for a database of 1018 images that are representative in variety for a much larger (commercial) database. From the features, a total of 504 aspects were derived.

5.2. Experimental Results

Fig. 3 shows precision-recall graphs based on the ranking of the target images for the scenarios outlined above based on 1000 simulations for each scenario. For the simulations random target aspects were sampled such that the number of target images was at least 10, and additionally the selected aspects corresponded to different features. Only salient aspects with p_{db} value of at most 0.1 were considered as target aspects, and for each aspect 4 example images were selected. Several variations to the experiments have been performed (e.g. taking more example images, increasing the saliency threshold, or leaving out various feature groups). All showed the same relative performance of the four methods.

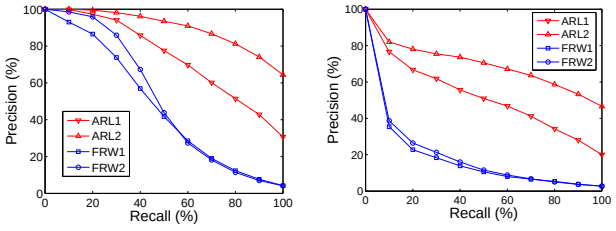


Figure 3. Target image precision-recall graphs obtained by methods ARL1, ARL2, FRW1 and FRW2, for the (a) full relevance scenario and (b) partial relevance scenario.

In the first scenario the aspect-based relevance feedback methods outperform the feature re-weighting methods mainly because the re-weighting methods assign too many features with high weights, which leads to poor precision. Similarly, $ARL2$ outperforms $ARL1$ as it selects fewer false positive aspects. In our retrieval system improved performance is obtained by using the adaptive threshold strategy of section 4.1 which represents a mix of ARL1 and ARL2. There, ARL1 is used in the initial stage where example set quality is expected to be low. We also reiterate the potential of using explicit user interaction to decide on which p-values to use. In scenario II the performance of the

feature re-weighting methods further deteriorates due to additional difficulty in selecting the correct features as clustering is less clear due to the partial relevance of the examples.

Fig. 4 demonstrates that the ARL methods lead to a higher accumulation of target aspects in the top ranking images. For each target aspect the fraction of images having that aspect was determined for the 50 highest ranking images (corresponding to the first selection display in our retrieval system). Shown are the average fractions over the target aspects. Note that the high accumulation of the ARL methods is preserved under the partial relevance scenario.

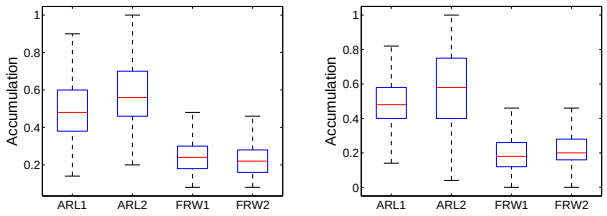


Figure 4. Accumulation box-plots for the (a) full relevance scenario, and (b) partial relevance scenario.

6. Conclusion

The aspect-based relevance learning method guarantees that feedback on feature values is accepted only once evidential support that the feedback was intended is sufficiently strong. This is a beneficial property for retrieval applications where example selection by partial relevance is important, as for instance in our retrieval system for decoration designs.

Our experience with the retrieval system and first simulation results confirm the feasibility of the approach. When features are reliable, generally few positive examples are required, and there is a regular progression to the target class without needing to browse through many selection displays in search for suitable examples. Another interesting property is that there is no need for negative examples solely for obtaining sufficient data for classification.

Future work will be directed at detailed comparison to other relevance learning methods. Also we intend to study generalizations such as fuzzy aspect possession, and alternative relevance ranking schemes.

References

Ciocca, G., & Schettini, R. (1999). A relevance feedback mechanism for content-based image retrieval. *Information Proc. and Management*, *35*, 605–632.

Huiskes, M. (2005a). Aspect-based relevance learning for image retrieval. In W. Leow (Ed.), *Proceedings of the international conference on image and video retrieval (CIVR), LNCS 3568*, 639–649. Springer.

Huiskes, M. (2005b). Indexing, learning and content-based retrieval for special purpose image databases. In M. Zelkowich (Ed.), *Advances in computers*. Elsevier.

MacArthur, S., Brodley, C., & Shyu, C. (2000). Relevance feedback decision trees in content-based image retrieval. *IEEE CBAIVL*, 68–72.

Peng, J., Bhanu, B., & Qing, S. (1999). Probabilistic feature relevance learning for content-based image retrieval. *Comp. Vis. and Image Und.*, *75*, 150–164.

Rocchio, J. (1971). Relevance feedback in information retrieval. In G. Salton (Ed.), *The SMART retrieval system: experiments in automatic document processing*, 313–323. Prentice-Hall.

Rui, Y., Huang, T., Ortega, M., & Mehrotra, S. (1998). Relevance feedback: A power tool for interactive content-based image retrieval. *IEEE Trans. Circ. and Syst. for Video Technology*, *8*, 644–655.

Salton, G., & Buckley, C. (1990). Improving retrieval performance by relevance feedback. *J. Am. Soc. Inf. Sci.*, *41*, 287–288.

Squire, D., Müller, W., Müller, H., & Raki, J. (1999). Content-based query of image databases, inspirations from text retrieval. *SCIA'99*, 143–149.

Tieu, K., & Viola, P. (2004). Boosting image retrieval. *Int. Journal of Computer Vision*, *56*, 17–36.

Tong, S., & Chang, E. (2001). Support vector machine active learning for image retrieval. *Proc. of 9th ACM Int'l Conf. on Multimedia*, 107–118.

Wu, P., & Manjunath, B. (2001). Adaptive NN search for relevance feedback in large image databases. *Proc. of 9th ACM Int'l Conf. on MM*, 89–97.

Zhang, H., Zheng, C., Li, M., & Su, Z. (2003). Relevance feedback and learning in content-based image search. *WWW: Internet and web information systems*, *6*, 131–155.

Zhou, X., & Huang, T. (2003). Relevance feedback in image retrieval: a comprehensive review. *ACM Multimedia Systems Journal*, *8*, 536–544.

Blotch Detection in Archive Film Restoration by Adaptive Learning

Attila Licsár LICSARA@ALMOS.VEIN.HU
Department of Image Processing and Neurocomputing, University of Veszprém, Veszprém, Hungary

Tamás Szirányi SZIRANYI@SZTAKI.HU
Analogical & Neural Computing Laboratory, Hungarian Academy of Sciences, Budapest, Hungary

László Czúni CZUNI@ALMOS.VEIN.HU
Department of Image Processing and Neurocomputing, University of Veszprém, Veszprém, Hungary

Abstract

We have developed a new semi-automatic blotch detection method with an object based post-processing to reduce false alarm results on archive films. Blotches can be modeled as temporal discontinuities of pixel intensity not originating from object motion, occlusion, disocclusion or non-rigid objects. In practice, usually, after the automatic detection step an operator, significantly decreasing the efficiency of the restoration process, removes the false alarms manually. Our proposed method reduces this manual intervention by a trainable classification method that filters out most of the false detection results. The examined classification methods are evaluated on ground truth data sets generated from real archive sequences. In our new evaluation method, we measure the number of the erroneously classified blotch objects describing the accuracy of the classification methods.

1. Introduction

National film archives store huge amounts of degraded films to be restored. Digital restoration methods can provide semi-automatic processing that results in efficient and cost effective saving and reconstruction of the film heritage. This should be achieved by fast, robust and automatic processing with a minimal human intervention, which is the bottleneck in the restoration work. Archive films suffer from several degradations such as blotches, scratches, flickering (intensity fluctuation), image vibration (displacement of adjacent frames), fading, discoloring, etc. One-frame defects are typical film artifacts, which are mostly visible as blotches. These artifacts appear at random positions on consecutive frames and have arbitrary shape, size and varying range of intensity (from bright to dark). Blotches are usually caused by dirt, damage of the film surface and chemical or biological processes such as mold. One-frame defects can be modeled as temporal intensity discontinuities, hence false detection results originate from object motion (e.g. occlusion), non-rigid objects or erroneous motion estimation. A typical restoration procedure of one-frame defects is the following: (1) detection of the defected regions, (2) interpolation of the corrupt image regions by spatio-temporal inpainting methods. In this paper, we focused only on the detection of artifacts. Automatic blotch detection methods could generate false alarms (non-valid blotches classified as valid blotches), and the inpainting of them causes loss of original image details not acceptable by archivists. Our concept is detection with lower false alarm rate rather than with maximal detection rate. In practice, after the automatic detection an operator manually verifies and classifies candidates as false alarms or as valid artifacts. Hence, an automatic method is needed to reduce false alarms by classifying the previously detected blotch candidates. Our paper introduces an object based classification method that minimizes the time-consuming manual correction steps. The proposed method, after trained by an operator, automatically filters out some of the detected blotches to avoid the further processing of the image based on false alarm detections.

2. Blotch Detection and Post-Processing Methods

Blotch detection methods can be divided into two groups: (1) spatial detection by analysis of the contrast or local maxima/minima; (2) spatio-temporal methods based on the detection of temporal discontinuities. The first group includes morphological operator based methods (Naranjo, 2004; Joyeux, 2001; Tenze, 2000) resulting in low complexity because they do not require temporal analysis such as motion estimation. Methods in the second group are based on the detection of temporal discontinuities like

Appearing in Proceedings of the workshop Machine Learning Techniques for Processing Multimedia Content, Bonn, Germany, 2005.

the SDI (Spike Detection Index) (Kokaram, 1998), ROD (Rank Ordered Differences) (Nadenau, 1997; Gangal, 2004), MRF (Markov Random Field) (Kokaram, 1995) methods, simplified ROD detector (SROD) (Roosmalen, 1999). According to the comparative evaluation of these methods (Nadenau, 1997), ROD resulted in the best trade-off between accuracy and computational cost. Post-processing and pre-processing methods of detection methods are intended to reduce false alarms preserving detection rate. Post-processing methods are hysteresis thresholding or constrained region growing (Roosmalen, 1999) of blotches to correct partially detected blotches and to reduce false alarms. A typical pre-processing step before detection is the bi-directional motion compensation of neighboring frames, which reduces false alarms due to the local/global motion. Important factor is the motion estimation complexity because we are processing a huge amount of data of high-resolution sequences (2000x1500 pixels (2K) or 6K resolution (Czúni, 2004)) of 35mm archive films. Others usually apply block matching methods with multi-scale processing (Buisson, 2003; Roosmalen, 1999; Nadenau, 1997), and several heuristic (e.g. logarithmic) searching methods (Gangal, 2004; Buisson, 2003) in order to increase the computation speed. The drawbacks are that they do not guarantee optimal detection and the estimation gives only pixel accuracy.

3. New Blotch Detection Method with Object Based Post-Processing

Our aim is to decrease the rate of false detections and to accelerate the motion estimation. The main steps of the algorithm are the following (the detailed description of the algorithm can be found in (Licsár, 2005)):

1) SROD method analyses the maximal intensity difference between the actual pixels and the set of neighboring pixels on the preceding and consecutive images and then segments defected points according to the blotch detection parameter.

2) The mask, obtained by the previous step, involves false alarms mostly due to the scene motion. In our new concept, the motion estimation and compensation is a post-processing method and it is only computed on the previously detected mask pixels. We use a hierarchical gradient-based motion estimation method (Bergen, 1990) that is based on Horn and Schunck's (Horn, 1981) optical flow constraint equations where the motion is modeled by simple translations. This approach significantly reduces the computational complexity of motion compensation (MC-SROD).

3) The next step is the object based blotch classification of image features extracted from the image intensities (luminance channel) given by the actual blotch mask. This step reduces the residual false alarms analyzing only the current frame because motion compensation does not give satisfactory result due to the incorrect motion estimation in regions with complex motion, non-rigid motion, occlusion, disocclusion, motion blur etc. Extracted features are as follows: maximal horizontal and vertical intensity change inside the blotch area, local internal/external intensity contrast inside/outside the blotch area, local internal/external mean and variance of the blotch area, and perimeter of the blotch.

4. Performance Analysis of the Object Based Classification

The object based classification is carried out by a feed-forward neural network (NN) or by a support vector machine (SVM). The configuration of the NN was 8 input features and 1 output result with two hidden layers (with 12 and 10 neurons), the training algorithm was the back-propagation method. The SVM method uses radial basis function kernel (RBF), where the gamma was 0.125. Usually, the performance of detection methods is statistically evaluated on artificially generated test sequences where randomly selected image regions are replaced with a rendered blotch, defined by a simple blotch model, i.e. homogenous blotch with sharp contour. On the contrary, we made our tests on real archive sequences where the ground truth data set of real artifacts is produced by manually marking blotches on the digitized films. This is important because our post-processing method analyses the intensity information of the detected regions and in case of artificially generated blotches the applied blotch model determines these. This process is semi-automatic because the operator frames each blotch by a rectangle and this region will be automatically replaced with the help of neighboring frames. The difference image between original and replaced region determines the blotch mask by a threshold based segmentation. If the result is not satisfactory, a manually painted mask can determine the one-frame defect. During the statistical evaluation the detected smaller blotches, for which the perimeter is not greater than 10, are skipped owing to the huge amount of image noise (e.g. grain noise). The test sequences are from the first Hungarian color film "Mattie the Goose-Boy" at 2K resolution. Our 5 test sequences include 50 frames with the corresponding blotch masks. This ground truth database involves varied content such as local/global motion with complex motion and/or motion blur or zero camera motion with small amount of local motion. The main questions are how the false alarm and detection rates and the amount of manual interventions change thanks to the automatic blotch classification.

4.1 Experimental Settings

Our supervised training is divided into two phases: (1) detection and displaying of the blotches by the previously described automatic method in the selected frame for training purposes; (2) operator selects typical examples of

positive (detected and displayed blotch is a valid artifact) and negative (non-valid artifact) samples by selection or framing them with a rectangle. The initial detection of the blotch mask is computed only once so the training and the classification steps have low computational cost (typically classification takes less than 1 second on a 2K frame). After the training process, the blotches are detected automatically. If the result is not satisfactory, the operator can expand the training set with new samples or rebuild the whole training set.

4.2 Performance Results

In our first test, we examined the changing of the false alarm (FA) and detection rate (DR) carrying out our object based post-processing method. We measured the FA and DR values without the object based classification (MC-SROD) and with NN and SVM classification methods (NN-MC-SROD, SVM-MC-SROD). The reduction ratios (RR) of DR and FA are calculated with the proportion of rates after and before the blotch classification step as the follows:

$$RR_{[NN,SVM]}^{[DR,FA]} = \frac{[DR,FA]_{[NN,SVM]-MC-SROD}}{[DR,FA]_{MC-SROD}} \quad (1)$$

The frames of the first training set (Set I.) were collected from our 5 sequences. The set involves 160 positive and negative samples. The second training set (Set II.) includes the Set I. and plus 30 positive and negative samples collected from the actual sequence. We summarized the reduction ratios after the NN and SVM based classification in Table 1. The DR reduction ratio of 0.942 means that DR has reduced with about 6 % after the NN based blotch classification. It can be seen that the accuracy could be increased if the operator makes an additional training before the restoration of each sequence but the initial training set is more essential.

Table 1. Reduction ratio of the average false alarm (FA) and detection rate (DR) after NN and SVM based object classification with two training set.

Training set	*NN method*		*SVM method*	
	Reduction ratios (RR)			
	FA	DR	FA	DR
Set I.	0.158	0.942	0.109	0.882
Set II.	0.143	0.955	0.108	0.894

In our second experiment, we measured that, after the automatic classification method, how many manual reclassification (from valid blotch to non-valid blotch or vice versa) would be necessary to achieve the best matching with the ground truth database (Figure 1). Since the time consumption of the manual correction step is proportional to the number of reclassified blotches, this value can describe the amount of manual intervention.

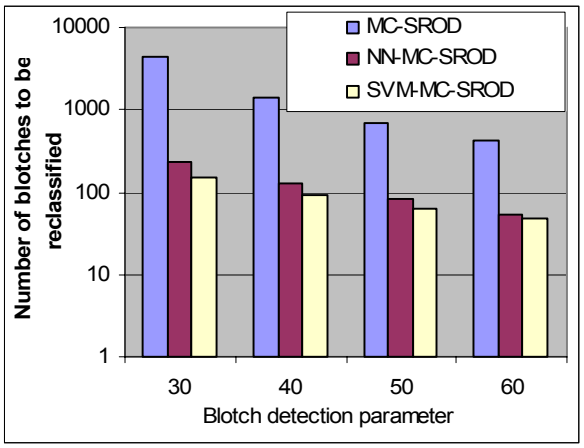


Figure 1. Number of the reclassified blotches to approximate the ground truth data set, where the blotch mask was generated with different methods (horizontal axes is in logarithmic scale).

We found that NN-MC-SROD and SVM-MC-SROD reduced significantly the number of blotches to be manualy reclassified. The reduction of interventions is between 80 – 95 %. If we examine the number of false negative results (Table 2) of NN and SVM classification methods (false negative: valid blotches classified as non-valid blotches), it can be experienced that the operator should make less correction steps to maximize the detection rate with the NN classification method. Therefore, the operator needs much less manual work to achieve higher detection rate. Otherwise, if the main criterion is the low false alarm without any human intervention (fully automatic mode) the SVM based method is favorable due to the lower false alarm rate.

Table 2. Number of false negative (valid blotches classified as non-valid blotches) and false alarm detected objects after the object based classification methods measuring at 50 frames.

Blotch detection parameter	*NN-MC-SROD*		*SVM-MC-SROD*	
	False negative	False alarm	False negative	False alarm
30	24	204	49	101
40	8	122	26	64
50	8	74	18	45
60	6	46	14	33

In Figure 2 there is an example of our classification method where circles indicate the detected blotches and

the radius of circles illustrate the size of the blotches (in case of smaller circles blotches might not be seen at this resolution). Red circles indicate that our classification method identified the blotch as non-valid blotch otherwise it is recognized as valid blotch.

Figure 2. Classification results of our post-processing method: original image (top) and the results of the object based classification method (bottom), where false alarms of the initial detections are marked with red circle.

5. Conclusion

We showed that the proposed object based classification method significantly improves the detection efficiency by the automatic reduction of the false alarms and minimizes the necessity of the human intervention.

Acknowledgments

Thanks to the Hungarian Film Archive for the scanned films. The project was supported by the Hungarian Ministry of Economy and Transport and the European Union in the frameworks of the ECOP/GVOP AKF388 and the NoE MUSCLE project of Eu.

References

Bergen, J. R. & Hingorani, R. (1990) Hierarchical Motion-Based Frame Rate Conversion. (Technical Report). David Sarnoff Research Center, Princeton.

Buisson, O., Boukir, S. & Besserer. B. (2003) Motion compensated film restoration. *Machine Vision and Applications*. Springer-Verlag, 13, 201-212.

Czúni, L. at al. (2004) Digital Motion Picture Restoration System for Film Archives (DIMORF). *SMPTE Motion Imaging Journal*. 113, 170-176.

Gangal, A. et al. (2004) An improved motion-compensated restoration method for damage color motion picture film. *Signal Proc.: Image Communication,* 19, 353-368.

Horn, B. K. P. & Schunk, B. C. (1981) Determining Optical Flow. *Artificial Intelligence*, 17, 185-203.

Joyeux, L., Boukir, S., Besserer, B. & Buisson, O. (2001) Reconstruction of degraded image sequences. Application to film restoration, *IMAVIS*, 19, 503-516.

Kokaram, A. C., Morris, R., Fitzgerald, W. & Rayner, P. (1995) Detection of missing data in image sequences. *IEEE Image Processing*, 1496-1508.

Kokaram. A. C., (1998) *Motion Picture Restoration: Digital Algorithms for Artefact Suppression in Degraded Motion Picture Film and Video*, Springer Verlag.

Licsár, A., Czúni, L. & Szirányi, T. (2005) Trainable post-processing method to reduce false alarms in the detection of small blotches of archive films. Proc. Intern. Conf. on Image Processing (ICIP), Genova.

Nadenau, M. J. & Mitra, S. K. (1997) Blotch and scratch detection in image sequences based on rank ordered differences. *Time-Varying Image Processing and Moving Object Recognition*, Elsevier.

Naranjo, V. at al. (2004) Morphological Lambda Reconstruction applied to restoration of blotches in old films. Proc. of the 4th IASTED Intern. Conf. on Visualisation, Imaging and Image Processing, Spain.

Roosmalen, P. M. B., Biemond, J. & Lagendijk, R. L. (1999) Restoration and storage of film and video archive material, *Signal Processing for Multimedia*.

Tenze, L., Ramponi, G. & Carrato, S. (2000) Blotches correction and contrast enhancement for old film pictures. Proc. Intern. Conf. on Image Processing (ICIP), Canada, 660-663.

Music Classification with Partial Selection Based on Confidence Measures

Wei Chai CHAIWEI@MEDIA.MIT.EDU
Barry Vercoe BV@MEDIA.MIT.EDU
Media Laboratory, Massachusetts Institute of Technology, 20 Ames Street, Cambridge, MA 02139 USA

Abstract

Music classification is a useful technique that enables automation of labeling musical data for searching and browsing. One method for music classification is to label the sequence based on the labels of individual frames. This paper investigates the performance of using confidence measures to select only the most "useful" frames to make the decision of the whole sequence. Confidence measures for Support Vector Machines (SVM) and Predictive Automatic Relevance Determination by Expectation-propagation (Pred-ARD-EP) are particularly examined. Experimental result shows that selecting frames based on confidence significantly outperform selecting frames randomly and the confidence measures do, to some extent, capture the "usefulness" of musical parts for classification.

1. Introduction

With the tremendous growth of digital music on computers, personal electronics and the Internet, music information retrieval has become a rapidly emerging research field. Music classification is one of the popular topics in this field, which enables automation of labeling musical data for searching and browsing.

Methods for music classification can be summarized into two categories. The first method is to segment the musical signal into frames, classify each frame independently, and then assign the sequence to the class to which most of the frames belong. It can be regarded as using multiple classifiers to vote for the label of the whole sequence. This technique works fairly well for timbre-related classifications. Pye (2000) and Tzanetakis (2002) studied genre classification. Whitman (2001), Berenzweig (2001, 2002) and Kim (2002) investigated artist/singer classification. In addition to this frame-based classification framework, the second method attempted to use features of the whole sequence (e.g., emotion detection by Liu, 2003), or use models capturing the dynamic of the sequence (e.g., Explicit Time Modeling with Neural Network and Hidden Markov Models for genre classification by Soltau, 1998) for music classification.

This paper focuses on the first method for music classification, investigating the relative usefulness of different musical parts when making the final decision of the whole musical piece, though the same idea might also be explored for the second method.

If humans are asked to listen to a piece of music and tell who is the singer or who is the composer, we typically will hold our decision until we get to a specific point which can show the characteristics of that singer or composer in our mind (called the *signature* of the artist). Thus, the question that this paper addresses is which part of a piece contributes most to a judgment about music's category when applying the first classification framework and whether what is "important" for machines (measured by *confidence*) is consistent with human intuition.

This paper will explore two classification techniques (Support Vector Machines and Predictive Automatic Relevance Determination by Expectation-propagation) and their confidence measures, and see whether we can throw away the "noisy" frames and use only the "informative" frames to achieve equally good or better classification performance.

This is similar to Berenzweig's method (2002), which tried to first locate the vocal part of musical signals and use only the vocal part to improve the accuracy of singer identification. The main difference is that here the algorithm does not assume any prior knowledge about which parts are "informative" (e.g., the vocal part is more informative than the accompaniment part for singer identification); on the contrary, we let the classifier itself choose the most "informative" parts by having been given a proper confidence measure. We then can analyze whether the algorithmically chosen parts are consistent with our prior knowledge. Therefore, to some extent, it is a reverse problem of Berenzweig's: if we can find a proper confidence measure, the algorithm should choose the vocal parts automatically for singer identification.

The remainder of this paper is organized as follows. Section 2 introduces the framework of music

Appearing in *Proceedings of the workshop Machine Learning Techniques for Processing Multimedia Content*, Bonn, Germany, 2005.

classification, the two classifiers (SVM and Pred-ARD-EP) and their confidence measures. Section 3 presents the experiments and results. Section 4 concludes the paper and proposes some future work.

2. Approach

2.1 Procedure of Music Classification

The first three steps are the same as the most-widely used approach for music classification:

1. Segment the signal into frames and compute the feature of each frame (e.g., FFT, Mel-Frequency Cepstral Coefficients);
2. Train a classifier using all the frames of the training signals independently;
3. Given a test signal, apply the classifier to the frames of the sequence and assign it to the class to which most of the frames belong;

Following these is one additional step:

4. Instead of using all the frames of a test signal for determining its label, a portion of the frames are selected according to a specific rule (e.g., select randomly, select the ones with the highest confidence) to determine the label of the piece.

Again, the last step can be regarded as choosing from a collection of classifiers for the final judgment. Thus, if we select frames based on confidence, the confidence measure should be able to capture the reliability of the classification, i.e., how certain that the classification is correct.

2.2 Classifiers and Confidence Measures

Let us consider discriminative models for classification. Suppose the discriminant function $S(\mathbf{x}) = \hat{y}$ is obtained by training a classifier, the confidence of classifying a test sample should be the predictive posterior distribution:

$$C(\mathbf{x}) = P(y = \hat{y} \mid \mathbf{x}) = P(y = S(\mathbf{x}) \mid \mathbf{x}) \quad (1)$$

However, it is generally not easy to estimate the posterior distribution. Thus, we need a way to estimate it, which is natural for some types of classifiers, while not so natural for some others.

In the following, we will focus on linear classification, i.e., $S(\mathbf{x}) = \hat{y} = sign(\mathbf{w}^T\mathbf{x})$, since nonlinearity can easily be incorporated by kernelizing the input point. Among the linear classifiers, Support Vector Machines (SVM) is a representative of non-Bayesian approach, while Bayes Point Machine (BPM) is a representative of Bayesian approach. This paper will investigate these two linear classifiers and their corresponding confidence measures.

For BPM, **w** is modeled as a random vector instead of an unknown parameter vector. Estimating the posterior distribution for BPM was extensively investigated by Minka (2001) and Qi (2002; 2004). Here Predictive Automatic Relevance Determination by Expectation-propagation (Pred-ARD-EP), an iterative algorithm to compute an approximate posterior distribution, will be used for estimating the predictive posterior distribution:

$$\begin{aligned} C(\mathbf{x}) &= P(y = \hat{y} \mid \mathbf{x}, D) \\ &= \int_{\mathbf{w}} P(\hat{y} \mid \mathbf{x}, \mathbf{w})\, p(\mathbf{w} \mid D) d\mathbf{w} = \Psi(z) \end{aligned} \quad (2)$$

$$z = \frac{(\hat{y}\mathbf{m}_w)^T \mathbf{x}}{\sqrt{\mathbf{x}^T \mathbf{V}_w \mathbf{x}}} \quad (3)$$

where D is the training set, **x** is the kernelized input point, $\hat{y}$ is the predictive label of **x**. $\Psi(a)$ can be a step function, i.e., $\Psi(a) = 1$ if $a > 0$ and $\Psi(a) = 0$ if $a \le 0$. We can also use the logistic function or probit model as $\Psi(\cdot)$. $\mathbf{m}_w$ and $\mathbf{V}_w$ are mean and covariance matrix of the posterior distribution of **w**, i.e., $p(\mathbf{w} \mid t, a) = N(\mathbf{m}_w, \mathbf{V}_w)$. a is a hyper-parameter vector in the prior of **w**, i.e., $p(\mathbf{w} \mid a) = N(0, \mathrm{diag}(a))$.

Estimating the posterior distribution for SVM might not be very intuitive, because the idea for SVM is to maximize the margin instead of estimating the posterior distribution. If we mimic the confidence measure for BPM, we obtain

$$C(\mathbf{x}) = \Psi(z) \quad (4)$$

$$z = (\hat{y}\mathbf{w})^T \mathbf{x} \quad (5)$$

Thus, the confidence measure for Pred-ARD-EP is similar to that for SVM except that it is normalized by the square root of the covariance projected on the data point. The confidence measure for SVM is proportional to the distance between the input point and the classification boundary.

2.3 Features and Parameters

For both SVM and Pred-ARD-EP, RBF basis function (Eq. 6) was used with $s = 5$. Probit model was used as $\Psi(\cdot)$. The maximum lagrangian value in SVM (i.e., C) was set to 30. All the parameters were tuned based on several trials.

$$K(x, x') = \exp\left(-\frac{\| x - x' \|^2}{2s^2}\right) \quad (6)$$

The feature used for both experiments was Mel-frequency Cepstral Coefficients (MFCCs). It is widely used for speech and audio signals.

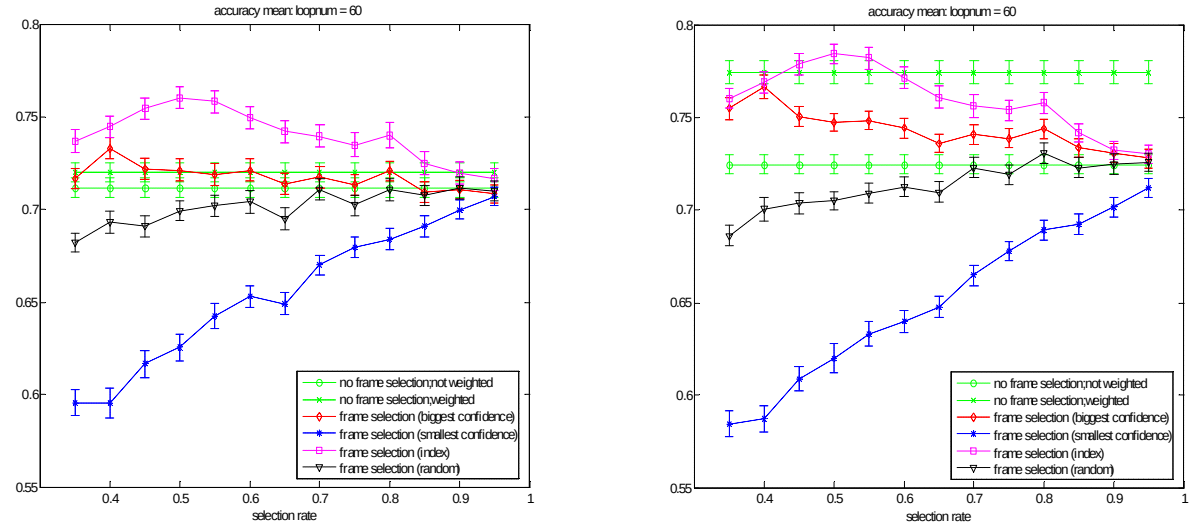


Figure 1. Accuracy of Genre Classification with Noise (left: Pred-ARD-EP; right: SVM)

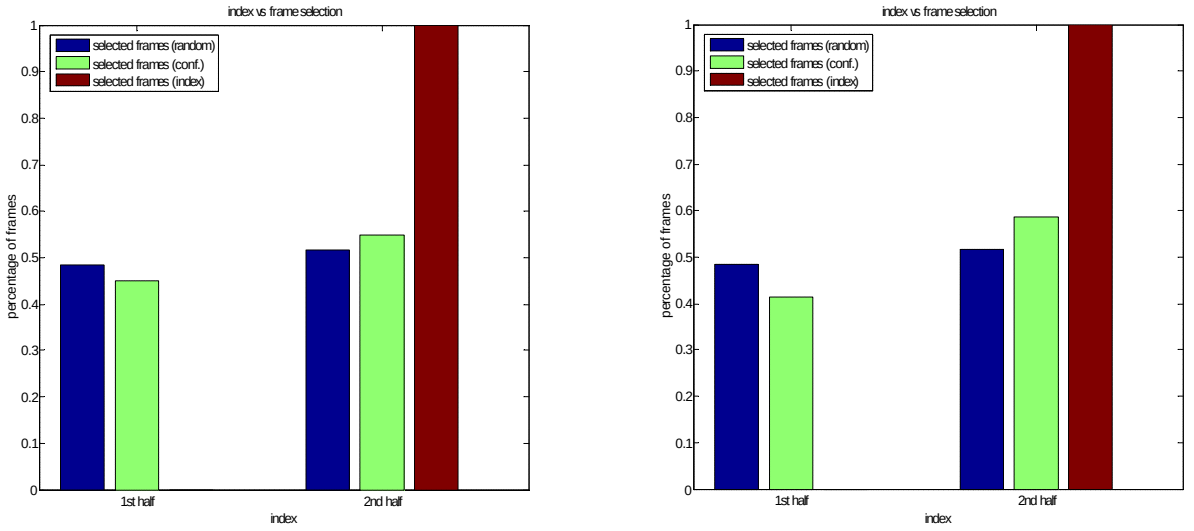


Figure 2. Index distribution of selected frames at selection rate 50% (left: Pred-ARD-EP; right: SVM)

3. Experiments and Results

Two data sets were chosen for the convenience of analyzing the correlation between algorithmically selected frames based on confidence and intuitively selected frames base on prior knowledge. The first is a genre classification data set with the first half of each sequence replaced by white noise. The second is a data set of monophonic singing voice for gender classification. In both cases, we only consider binary classifications.

Specifically, for either experiment, the data set was sampled at 11kHz sampling rate. Analysis was performed using frame size of 450 samples (~40 msec) and frames were taken every 225 samples (~20 msec). MFCCs were computed for each frame. Only every 25th data frame was used for training and testing because of the computer memory constraint. 30 % of the sequences were used for training, while 70% were used for testing. The performance was averaged over 60 trials.

3.1 Experiment 1: Genre Classification of Noisy Musical Signals

The data set used in this experiment consists of 112 orchestra recordings and 45 Jazz recordings of 10 seconds each. The MFCCs of the first half frames of each

sequence (both training and testing) were replaced by random noise normally distributed with mean and standard deviation of the original data.

The results from the experiment are summarized in Figure 1, which shows the percentages of sequences correctly classified.

In Figure 1, the x-axis denotes the selection rate, which denotes the percentage of frames selected according to some criterion. For example, selecting frames with highest confidence at a selection rate 60% means that the top 60% frames with the highest confidence will be counted for the final decision of the label of the whole sequence, while the other 40% frames will simply be ignored. The two horizontal lines are baselines, corresponding to the performances using all the frames available to each sequence (the above is confidence-weighted meaning each frame contributes differently to the label assignment of the whole sequence based on confidence; the below is not confidence-weighted). The other four curves, from top to the bottom, correspond to:

a. Selecting frames appearing later in the piece (thus, larger frame indexes and fewer noisy frames),

b. Selecting frames with highest confidence,

c. Selecting randomly,

d. Selecting frames with lowest confidence.

All these four curves approach the lower baseline when the selection rate goes to 1. It is easy to explain the peaks at selection rate 50% in curve *a*, since half of the frames were replaced by noise. The order of these four curves is consistent with our intuition. Curve *a* performed the best because it used the prior knowledge about data.

We also want to know the property of the selected frames. Figure 2 shows the percentage of selected frames (selecting by random, by confidence and by index) that are noise (first half of each piece) or not noise (second half of each piece) at selection rate 50%. As we expected, frame selection based on confidence does tend to select more frames at the second half of each piece (not entirely though).

Although this paper does not aim at comparing Pred-ARD-EP and SVM, for this data set, SVM outperformed Pred-ARD-EP. Here is one explanation of it. Due to the nature of the added noise with mean of all frames including both classes, most noisy samples fall between the two classes and thus near the classification boundary in SVM, so the confidence measure for SVM proportional to the distance between the data point and the boundary is a good estimate of confidence in this case. However, Pred-ARD-EP attempts to model the posterior distribution without considering that half of the data were actually noise and thus gets a worse performance and estimate of confidence.

3.2 Experiment 2: Gender Classification of Singing Voice

The data set used in this experiment consists of recordings of 45 male singers and 28 female singers, one sequence for each singer. All the other parameters are the same as the first experiment except that no noise was added to the data, since we here want to analyze whether the algorithmically selected frames are correlated with the vocal portion of the signal.

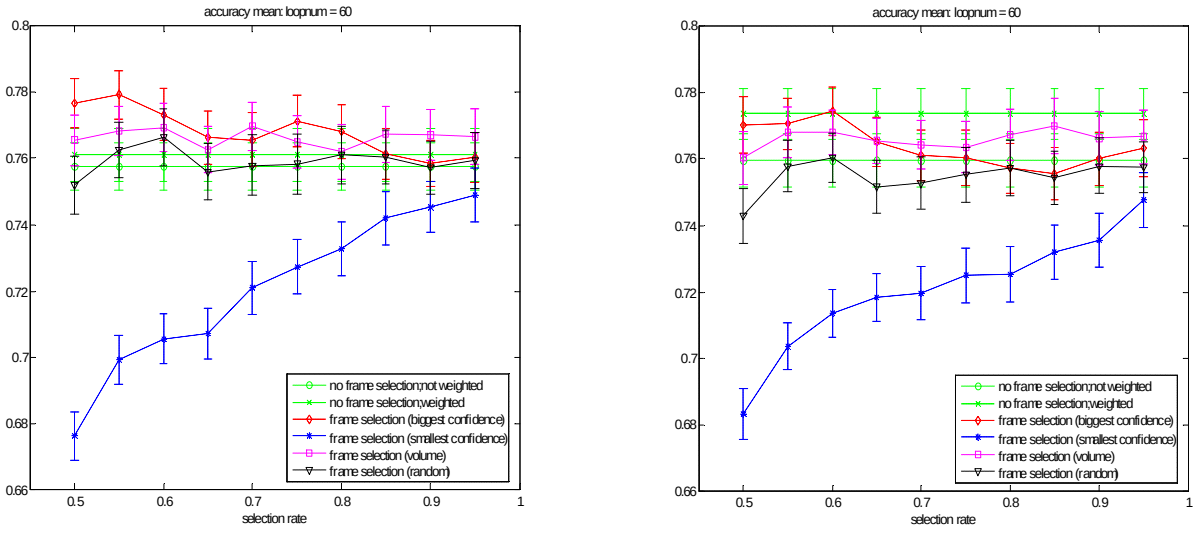


Figure 3. Accuracy of Gender Classification of Singing Voice (left: Pred-ARD-EP; right: SVM)

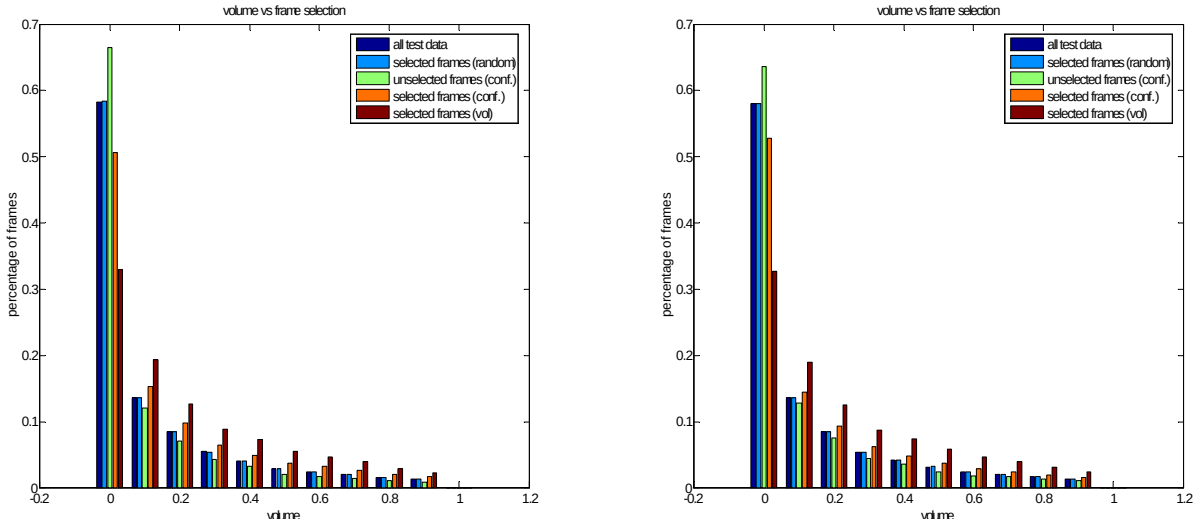


Figure 4. Volume distribution of selected frames at selection rate 55% (left: Pred-ARD-EP; right: SVM)

The results from the experiment are summarized in Figure 3. Similar to Figure 1, the two horizontal lines in Figure 3 are baselines. The other four curves, from top to the bottom, correspond to:

a. Selecting frames of the highest confidence,

b. Selecting frames of the highest energy,

c. Selecting randomly

d. Selecting frames of the lowest confidence.

In curve *b*, we used volume instead of index (i.e., location of the frame) as the criterion for selecting frames, because the data set consists of monophonic recording of singing voice and volume can be a good indicator of whether there is vocal at the time. The order of these four curves can be explained in the similar way as in the last experiment, except that, selecting frames based on prior knowledge seem not to outperform selecting frames based on confidence. The reason here is that volume itself cannot completely determine whether the frame contains vocal or not. For example, an environmental noise during recording can also cause high volume. It might be better to combine other features, e.g., pitch range, harmonicity, to determine the vocal parts.

Figure 4 shows the histogram (ten bins divided evenly from 0 to 1) of the volumes of selected frames at selection rate 55%. The five groups correspond to distributions of all test data, selected frames by random selection, discarded frames by confidence-based selection, selected frames by confidence-based selection and selected frames by volume-based selection. As we expected, the frame selection based on confidence does tend to select frames that are not silence.

To show the correlation between confidence selection and another vocal indicators – pitch range, Figure 5 shows a volume-pitch distribution difference between selected frames and unselected frames based on confidence. Pitch of each frame was estimated by autocorrelation. It clearly shows that the frame selection based on confidence tends to choose frames that have higher volume and pitches around 100~300Hz corresponding to the typical pitch range of human speakers. Note that most singers of the data set sang in a casual way. So, although the data set used here is singing voice instead of speech, the pitch range is not as high as typical professional singing.

4. Conclusion and Future Work

The experimental results demonstrate that the confidence measures do, to some extent, capture the importance of data, which is also consistent with the prior knowledge. The performance is at least equally good as the baseline (using all frames), slightly worse than using prior knowledge properly, but significantly better than selecting frames randomly. This is very similar to human perception: for humans to make a similar judgment (e.g., singer identification), given only the signature part should be as good as given the whole piece, while much better than given the trivial parts.

Although the classifiers tended to choose frames that are intuitively more "informative", they did not choose as many as they could: the noisy parts (in the first experiment) and the silent parts (in the second experiment) still seem to contribute to the classification. This should depend on how good the confidence measure is and how the classifier deals with noise. It suggests two directions in the future: exploring more confidence measure and investigating how different types of noise impact the estimate of confidence.

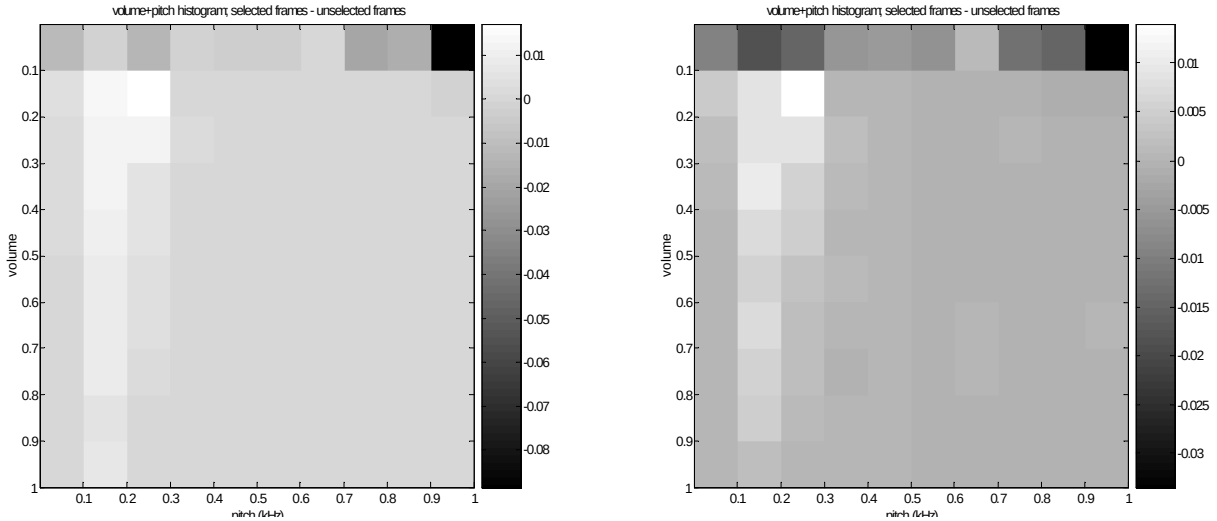


Figure 5. Pitch vs volume distribution of selected frames at selection rate 55% (left: Pred-ARD-EP; right: SVM)

Acknowledgments

This work was supported by the Digital Life consortium at the MIT Media Laboratory. I would like to especially thank Yuan Qi who gave me his Matlab code for Pred-ARD-EP.

References

Berenzweig, A. L. and Ellis, D. P. W. (2001). *Locating Singing Voice Segments within Music Signals*. Proceedings of Workshop on Applications of Signal Processing to Audio and Acoustics, Mohonk Mountain Resort, NY.

Berenzweig, A., Ellis, D., and Lawrence, S. (2002). *Using Voice Segments to Improve Artist Classification of Music*. Proceedings of International Conference on Virtual, Synthetic and Entertainment Audio.

Kim, Y. and Whitman, B. (2002). *Singer Identification in Popular Music Recordings Using Voice Coding Features.* In Proceedings of the 3rd International Conference on Music Information Retrieval. 13-17, Paris, France.

Liu, D., Lu, L., and Zhang, H.J. (2003). *Automatic mood detection from acoustic music data.* Proceedings of the International Conference on Music Information Retrieval. 13-17.

Minka, T. (2001). *A family of algorithms for approximate Bayesian inference.* PhD Thesis, Massachusetts Institute of Technology.

Pye, D. (2000). *Content-Based Methods for the Management of Digital Music.* Proceedings of International Conference on Acoustics, Speech, and Signal Processing.

Qi, Y., and Picard, R. W. (2002). *Context-sensitive Bayesian Classifiers and Application to Mouse Pressure Pattern Classification.* Proceedings of International Conference on Pattern Recognition, Québec City, Canada.

Qi, Y., Minka, T. P., Picard, R. W., and Ghahramani, Z. (2004). *Predictive Automatic Relevance Determination by Expectation Propagation.* Proceedings of International Conference on Machine Learning, Alberta, Canada.

Soltau, H., Schultz, T., and Westphal, M. (1998). *Recognition of Music Types*. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Seattle, WA. Piscataway, NJ.

Tzanetakis, G. M. (2002). *Analysis and Retrieval Systems for Audio Signals*. PhD Thesis, Computer Science Department, Princeton University.

Whitman, B., Flake, G. and Lawrence, S. (2001). *Artist Detection in Music with Minnowmatch*. In Proceedings of the IEEE Workshop on Neural Networks for Signal Processing, pp. 559-568. Falmouth, Massachusetts.

Interactive video retrieval based on multimodal dissimilarity representation

Eric Bruno ERIC.BRUNO@UNIGE.CH
Nicolas Moenne-Loccoz NICOLAS.MOENNE-LOCCOZ@UNIGE.CH
Stéphane Marchand-Maillet MARCHAND@UNIGE.CH
Viper group, Computer Vision and Multimedia Laboratory, University of Geneva
24 rue du Général Dufour, 1204 Geneva, Switzerland

Abstract

We present an approach to learn user semantic queries from dissimilarity representations of video audio-visual content. When dealing with large corpora of videos documents, using a feature-based representation calls for the online computation of distances between all documents and the query. Hence, a dissimilarity representation may be preferred because its offline computation speeds up the retrieval process. We show how distances related to visual and audio video features can directly be used to learn complex concepts from a set of positive and negative examples provided by the user. Based on the idea of dissimilarity spaces, we derive a low-dimensional multimodal representation space where an on-line and real-time classification is performed to learn user queries. The classification consists in maximizing a non-linear Fisher criterion to separate positive from negative examples. The evaluation, performed on the complete annotated TRECVid corpus, shows that our technique enables us to improve the precision of retrieval results.

1. Introduction

Determining semantic concepts by allowing users to iteratively refine their queries is a key issue in multimedia content-based retrieval. The relevance feedback loop allows to construct complex queries made out of positive and negative documents as examples. From this training set, a learning process should then extract relevant documents from feature spaces. Many relevance feedback techniques have been developed that operate directly in the feature space (Chang et al., 2003; Smith et al., 2003; Yan et al., 2003; Zhou & Huang, 2004).

Describing content of videos requires to deal in parallel with many high-dimensional feature spaces expressing the multimodal characteristics of the audiovisual stream. This mass of data makes retrieval operations computationally expensive when dealing directly with features. The simplest task of computing the distance between a query and all other elements becomes infeasible when involving tens of thousand of documents and thousand of feature space components. This problem is even more sensible when the similarity measures are complex functions or procedures, such as prediction functions for temporal distances (Bruno et al., 2005) or graph exploration for semantic similarities (Resnik, 1995).

A solution to allow on-line interaction would be to compute off-line monomodal dissimilarity relationships between elements and to use the dissimilarity matrices or distance-based indexing structures (Chávez et al., 2001) as an index for retrieval operations. The problem is then to find distance-based solutions that go beyond the classical k-NN approaches (Boldareva & Hiemstra, 2004) in order to perform effective classification and retrieval of semantic concepts. Pekalska *et al* (Pekalska et al., 2001) have proposed dissimilarity spaces where objects are represented not by their features but by their relative dissimilarities to a set of selected objects. These representations seem to form a convenient approach to tackle the similarity-based indexing and retrieval problem.

In this paper, we investigate the idea of dissimilarity spaces for the specific problem of multimedia document retrieval, and show how dissimilarities can be

Appearing in *Proceedings of the workshop Machine Learning Techniques for Processing Multimedia Content*, Bonn, Germany, 2005.

used to build a low-dimensional multimodal representation space where learning machines based on *eg* non-linear discriminant analysis could operate. Our thorough evaluation on the complete TRECVid corpus shows that this multimodal dissimilarity space allows to perform effective retrieval of video documents in real time, as defined in (Nielsen, 1993).

2. Classification in dissimilarity space

In the proposed retrieval system, video segments are represented by their dissimilarity relationships computed offline over several audiovisual features. The user can formulate complex queries by iteratively providing positive and negatives examples in an online relevance feedback loop. From this training data, the aim is to perform a real-time dissimilarity-based classification that will return relevant documents to user.

2.1. Dissimilarity space

Let $d(\mathbf{x}_i, \mathbf{x}_j)$ be the distance between elements i and j according to their descriptors $\mathbf{x} \in \mathcal{F}$. $\mathcal{F}$ expresses the (unavailable) original feature space. The dissimilarity space is defined as the mapping $\mathbf{d}(\mathbf{z}, \Omega) : \mathcal{F} \rightarrow \mathbb{R}^N$ given by (see (Pekalska et al., 2001) for details):

$$\mathbf{d}(\mathbf{z}, \Omega) = [d(\mathbf{z}, \mathbf{x}_1), d(\mathbf{z}, \mathbf{x}_2), \dots d(\mathbf{z}, \mathbf{x}_N)]. \quad (1)$$

The representation set $\Omega = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ is a subset of N objects defining the new space. The new "features" of an input element are now the dissimilarities between itself and the representation objects. As a consequence, learning or classification tools for feature representations are also directly available to deal with the dissimilarities.

The dimensionality of the dissimilarity space is directly linked to the size of Ω, which controls the approximation made on the original feature space (such an approximation could be computed using projection algorithms like classical scaling (Cox & Cox, 1995)). Increasing the number of elements in Ω increases the representation accuracy. On the other hand, we are interested in minimizing the space dimensionality so as to limit computation and to speed up the response time of the system. The selection of Ω will however be driven by considerations on the classification problem as explained now.

2.2. Non-linear discriminant analysis

Let us define the set T as the query formed out of positive and negative training examples (respectively denoted $\mathcal{P}$ and $\mathcal{N}$ with $T = \mathcal{P} \cup \mathcal{N}$), their coordinates in the dissimilarity space are respectively $\mathbf{d}_i^+ = \mathbf{d}(\mathbf{z}_{i \in \mathcal{P}}, \Omega)$ and $\mathbf{d}_i^- = \mathbf{d}(\mathbf{z}_{i \in \mathcal{N}}, \Omega)$.

Given a query T, the aim is therefore to find a relevance measure $D(\mathbf{d}_i) : \mathbb{R}^N \rightarrow \mathbb{R}$ that maximizes the following Fisher criterion

$$\max_D \frac{\sum_i D^2(\mathbf{d}_i^-)}{\sum_i D^2(\mathbf{d}_i^+)}. \quad (2)$$

The measure $D(\mathbf{d})$ gives us a new ranking function where positive elements tend to be placed at the top of the list while negatives one are pushed to the end.

Depending on the separability of the data according to a query T, the ranking function $D(\mathbf{d})$ may be chosen as a linear or non-linear function of the dissimilarities. Following the kernel machine formulation, $D(\mathbf{d})$ is written in both cases (linear or not) as an expansion of kernels centered on training patterns (Schölkopf & Smola, 2002):

$$D(\mathbf{d}) = \sum_{i \in T} \alpha_i k(\mathbf{d}, \mathbf{d}_i^{\pm}) + b. \quad (3)$$

Using such non-linear model in criterion (2) leads to the formulation of the Kernel Fisher Discriminant (KFD) (Mika et al., 1999). It has been shown that this problem can be solved by using mathematical programs (quadratic or linear). The proofs and the implementation of the algorithm we use to optimize (2) can be found in (Mika et al., 2000).

In general, we are dealing with a $1 + x$ class setup with 1 class associated to positives and x to negatives (Zhou & Huang, 2004). It is then needed to estimate complex decision functions to learn the semantic concepts, increasing the risk to encounter difficulties for choosing and tuning well-adapted kernels. However, selecting the representation set as the set of positive examples $\mathcal{P}$ turns the problem into a binary classification. Assuming that the positive examples are close to each other while all being far from negatives, the vectors $\mathbf{d}(\mathbf{z}_{i \in \mathcal{P}}, \mathcal{P})$ (*within* scatter) have norms lower than vectors $\mathbf{d}(\mathbf{z}_{i \in \mathcal{N}}, \mathcal{P})$ (*between* scatter), leading to a binarization of the classification, as illustrated in figure 1. In addition, this choice readily induces to work in a low dimensional space of $p = |\mathcal{P}|$ components, where online learning processes are dramatically speeded-up.

Kernel selection and setting is a critical issue to successfully learn queries. It actually decides upon the classical trade-off between over-fitting and generalization properties of the classifier and hence is very dependent of the considered dissimilarity space. This problem is discussed in the next section.

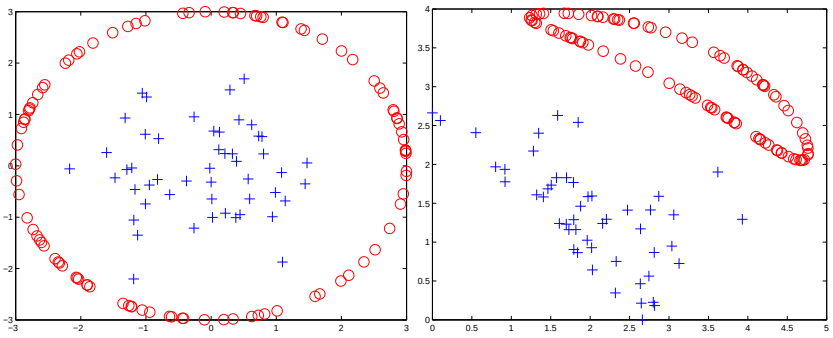


Figure 1. The $1 + x$ class problem in feature space (left) and dissimilarity space (right) where the representation objects are two points from the central class (cross)

3. Multimodal space

The video content is characterized by features corresponding to multiple modalities (*eg*, visual, audio, speech). Each of them leads to a dissimilarity matrix containing pairwise distances between all documents. Let us note d^{f_i} the distance measure applied on the feature space $\mathcal{F}_i$ and assume that dissimilarity matrices are known for M feature spaces. We define the multimodal dissimilarity space $\mathbf{d}$ as the concatenation of all monomodal spaces $\mathbf{d}^{f_i}$

$$\mathbf{d} = [\mathbf{d}^{f_1}, \mathbf{d}^{f_2}, \ldots, \mathbf{d}^{f_M}]. \tag{4}$$

The kernel function used in equation (3) now operates in a multimodal space. Its choice is then a critical issue to ensure the success of the modalities fusion coming from the resolution of equation (2). The RBF kernel $k(\mathbf{x}, \mathbf{y}) = e^{-(\mathbf{x}-\mathbf{y})^T \mathbf{A}(\mathbf{x}-\mathbf{y})}$ presents a convenient solution for our problem: it is indeed able to learn semantic concepts that are locally distributed within the representation space, and the scaling symmetric positive definite matrix $\mathbf{A}$ permits to tune the trade-off between over-fitting and generalization. As the input space is multimodal, the scaling matrix is constructed so as to allow independent scaling for each feature space, so that $\mathbf{A} = \text{diag}[\boldsymbol{\sigma}_{f_1}, \cdots, \boldsymbol{\sigma}_{f_M}]$. The vector $\boldsymbol{\sigma}_{f_i} \in \mathbb{R}^p$ is constant with all values equal to the scale parameter σ_{f_i} estimated for the dissimilarity space $\mathbf{d}^{f_i}$. Various approaches to automatically tune the scale parameters (Cristianini et al., 2001; Ong et al., 2003) have been proposed. However, the kernel estimation rely on an optimization of functionals that will drastically penalize the response time of the retrieval system. For this reason, the estimation of σ_{f_i} is based on a less optimal but simpler heuristic, adapting the model to the query

$$\sigma_{f_i} = C \cdot \text{median}_i(\min_j ||\mathbf{d}_i^+ - \mathbf{d}_j^-||^2). \tag{5}$$

In other words, the scale value in space $\mathbf{d}^{f_i}$ is set to be proportional to the median of all the minimum distances between the negative and the positive examples in that space. That way, the kernel becomes sharper as the two classes become closer to each other. The parameter C has been empirically set to 2.0.

4. Experimentations

Our multimodal interactive learning algorithm has been systematically experimented in the context of the video retrieval system we have developed. The segmented video documents, their multimodal description as well as manual annotations are stored in a database that keeps synchronized all data and allows large-scale evaluations of retrieval results.

The experimentation consists in making queries corresponding to annotated concepts and measuring the average precision (ratio of relevant documents in the retrieved list averaged over 50 queries) for retrieved lists of various lengths. The annotated positive examples are removed from the hitlist so that they are not taken into account when measuring the performance.

4.1. The video database

We use the complete annotated video corpus TRECVid-2003 composed of 133 hours of CNN and ABC news. Videos are segmented into shots and every shot has been annotated by several concepts. The speech transcripts extracted by Automatic Speech Recognition (ASR) at LIMSI laboratory (Gauvain et al., 2002) are also available.

We extracted the three following features from the 37'500 shots composing the corpus: Color histogram, Motion vector histogram and Word occurrence histogram (after stemming and stopping). The distance measures used are Euclidean for Color and Motion

histogram and intersection for Word occurrence histogram.

4.2. Results

We first test the validity of the monomodal dissimilarity space defined in section 2.2. We compare the precision of the retrieval when the classification is performed in the color feature space and in the corresponding dissimilarity space. Figure 2 shows results for two queries corresponding to two annotated concepts (*Basketball* and *Studio setting*). Whatever the size of the training set, the precision at the 100th position of the retrieval list is better when the dissimilarity space is used. It is important to note that the improvement becomes more important when the training set is small: when the class distributions to estimate are severely under-sampled (small training set), the simplification of the classification problem implied by the dissimilarity space (see section 2.2) is crucial for the success of the training stage.

We now evaluate how the combination of modalities may improve the retrieval efficiency. Figure 3 compares the average precision for several concepts when the query is learned in the monomodal spaces and in the multimodal space. We can observe that, even for queries where the raw features used are not well-suited (*Car* and *Desert*), the combination of the three modalities performs better than considering them separately. The precision graphs also compare the algorithm with a random retrieval (e.g seeking hits at random within the database). This comparison illustrates the capability of the algorithm to use low-level multimodal information to create models of semantic concepts defined by user. This improves drastically the performance of the search.

The following experiment tests how the retrieval precision evolves when the number of positive and negative documents grows. As figure 4 shows, the precision of the retrieval increases with the size of the training set until a point where adding more examples does not improve the performances anymore. This behavior illustrates how the users, by providing more and more examples (relevance feedback loop), can refine their queries until reaching the optimum of the classifier.

Finally, since we act in an interactive setup, we were interested in the computation time problem. The following measures (table 1) have been done on a PIV 2GHz and include the time to access the dissimilarity matrices (37500×37500), the building of the multimodal dissimilarity space and the training of the Fisher classifier. As the dimensionality of the representation space linearly depends on the number of positive examples, the response time increases according to their number. On the other hand, negative examples have less influence since they are just involved in the learning process.

Table 1. Response time

Neg. examples	20			100
Pos. examples	5	10	40	10
Resp. time (s)	1.4	2	7.4	4.3

5. Conclusion

We have presented a retrieval strategy for video documents. Based on a multimodal dissimilarity space associated to a non-linear discriminant analysis, the algorithm is able to take benefit from low-level multimodal descriptions of video documents and, as a consequence, to learn semantic queries from a limited number of input examples. The design of the dissimilarity space has been achieved so as to simplify the classification problem while building a low-dimensional representation of the data. The use of the positives examples as a representation set transforms the $1 + x$ setup into a binary classification problem. Sophisticated learning machines, such as the kernel Fisher discriminant analysis, can then directly be applied to classify data. As a result, semantic concepts are learned with more efficiency and queries on large databases are processed near real-time which authorizes the use of feedback loop as a search paradigm. Extensive evaluations on the TRECVid-2003 benchmark show the efficiency and the usability of the proposed multimodal space and fusion algorithm to retrieve documents within a large corpus of videos.

While the presented classification scheme has proved its value, the actual features considered to characterize the videos do not permit us to design a fully-capable and efficient video retrieval system. The design of new feature extractors related to new modalities (*e.g.* audio stream) and higher-level aspects of the content (*e.g.* face and object detection) is still a major issue. The addition of information sources should leads us to investigate more deeply the problems of the multimodal kernel design and setting as well as to determine the limits of the fusion scheme when a large number of features is used.

References

Boldareva, L., & Hiemstra, D. (2004). Interactive content-based retrieval using pre-computed object-object similarities. *Conference on Image and Video Retrieval, CIVR'04* (pp. 308–316). Dublin, Ireland.

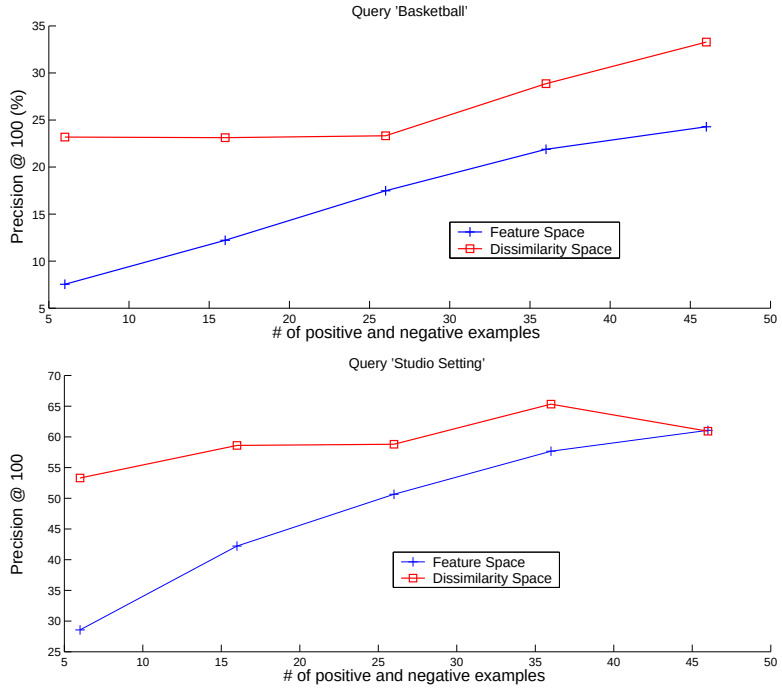


Figure 2. Average precision of the retrieval at 100 when the classification is performed directly in the color feature space (cross) and in the corresponding dissimilarity space (square) as the number of positive and negative examples increases.

Bruno, E., Moenne-Loccoz, N., & Marchand-Maillet, S. (2005). Unsupervised event discrimination based on non-linear temporal modelling of activity. *Pattern Analysis and Application, special issue on Video Event Mining*. (to appear).

Chang, E. Y., Li, B., Wu, G., & Go, K. (2003). Statistical learning for effective visual information retrieval. *Proceedings of the IEEE International Conference on Image Processing*.

Chávez, E., Navarro, G., Baeza-Yates, R., & Marroquin, J. (2001). Searching in metric spaces. *ACM Computing Surveys*, *33*, 273–321.

Cox, T., & Cox, M. (1995). *Multidimensional scaling*. London: Chapman & Hall.

Cristianini, N., Shawe-Taylor, J., Elisseeff, A., & Kandola, J. (2001). On kernel-target alignment. *Advances In Neural Information Processing Systems, Nips*.

Gauvain, J., Lamel, L., & Adda, G. (2002). The limsi broadcast news transcription system. *Speech Communication*, *37*, 89–108.

Mika, S., Rätsch, G., & Müller, K.-R. (2000). A mathematical programming approach to the kernel fisher algorithm. *NIPS* (pp. 591–597).

Mika, S., Rätsch, G., Weston, J., Schölkopf, B., & Müller, K.-R. (1999). Fisher discriminant analysis with kernels. *Neural Networks for Signal Processing IX* (pp. 41–48). IEEE.

Nielsen, J. (1993). *Usability engineering*. Boston, MA, USA: Academic Press.

Ong, C., Smola, A., & Williamson, R. (2003). Hyperkernels. *Advances in Neural Information Processing Systems, NIPS*.

Pekalska, E., Paclík, P., & Duin, R. (2001). A generalized kernel approach to dissimilarity-based classification. *Journal of Machine Learning Research*, *2*, 175–211.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. *14th International Joint Conference on Artificial Intelligence, IJCAI* (pp. 448–453). Montreal, Canada.

Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels*. MIT Press.

Smith, J. R., Jaimes, A., Lin, C.-Y., Naphade, M., Natsev, A., & Tseng, B. (2003). Interactive search fusion methods for video database retrieval. *IEEE International Conference on Image Processing (ICIP)*.

Yan, R., Hauptmann, A., & Jin, R. (2003). Negative pseudo-relevance feedback in content-based video retrieval. *Proceedings of ACM Multimedia (MM2003)*. Berkeley, USA.

Zhou, X., & Huang, T. (2004). Small sample learning during multimedia retrieval using biasmap. *Proceedings of the IEEE Conference on Pattern Recognition and Computer Vision, CVPR'01* (pp. 11–17). Hawaii.

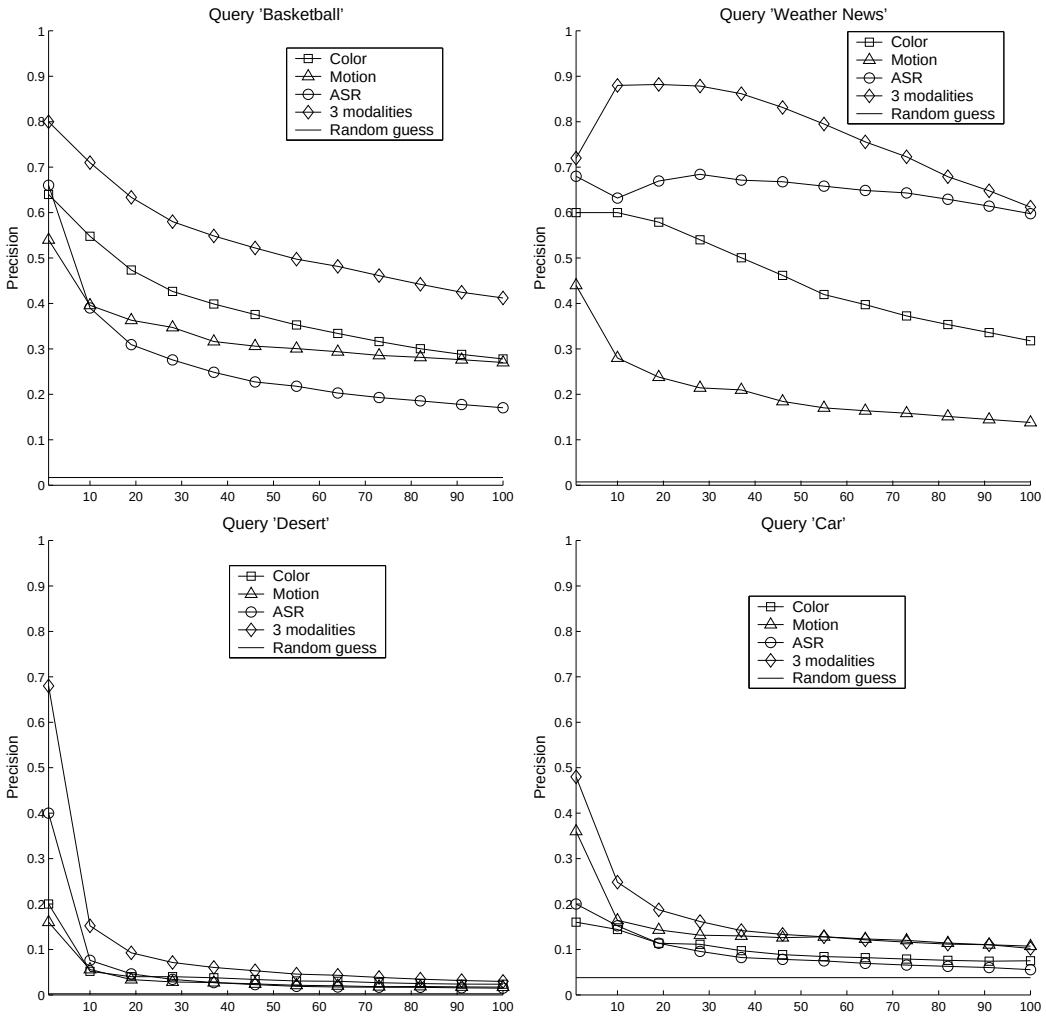


Figure 3. Average precision vs. length of retrieved lists for monomodal and multimodal dissimilarity spaces. The query is composed of 5 positive examples (annotated by the concept) and 20 negative examples randomly selected in the database. The "random guess" line is equal to the proportion of the concept in the database.

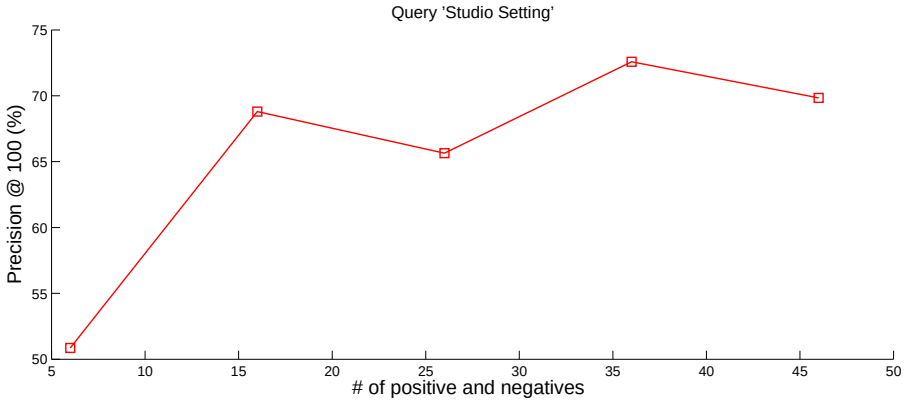


Figure 4. Average precision at 100 when positive examples and negative examples increase ($n_p = n_n$). Color, motion and ASR features are used.

Large margin multiple hyperplane classification for content-based multimedia retrieval

Serhiy Kosinov SERHIY.KOSINOV@CUI.UNIGE.CH
Ivan Titov IVAN.TITOV@CUI.UNIGE.CH
Stéphane Marchand-Maillet MARCHAND@CUI.UNIGE.CH
University of Geneva, 24 Rue du General-Dufour, CH-1211, Geneva, Switzerland

Abstract

This introductory note considers an application of content-based multimedia retrieval, where a semantic concept of a user query must be learned from only a few documents, provided as relevance feedback, that are vastly outnumbered by the irrelevant items of the collection. Formally, the problem in question is situated in the context of asymmetric classification where due to substantial imbalance, different classes are not treated equally. In contrast to the popular optimal separating hyperplane techniques that use only one hyperplane, an attempt is made to further exploit the asymmetric problem setting by incorporating multiple hyperplanes in a classifier so as to favor the under-represented class. Although the introduced modification leads to a more difficult optimization problem, a preliminary empirical evaluation of such a method in the asymmetric "one-against-all" classification setting provides encouraging results, which warrants further investigation.

1. Introduction

In this note, we consider the asymmetric classification problem setting, often encountered in content-based multimedia retrieval performed as a "one-against-all" classification scheme. The essence of the proposed technique is to increase the number of hyperplanes used in an optimal separating hyperplane classifier, so as to favor the under-represented class. Such a distinction that singles out a certain target class from the rest of the data, when modeled explicitly, has been previously shown to improve classification accuracy for undersampled and unbalanced data sets, (Akbani et al., 2004; Veropoulos et al., 1999; Zhou et al., 2004). While being applicable in the general classification scenario, the proposed method is designed to further exploit the asymmetry of the classification problem at hand.

The intuition behind the idea of introducing one or more extra hyperplanes in a classifier is exemplified in Figure 1, where it is shown how an additional hyperplane may improve the class separation margin, and thus have the potential to reduce the classification error rate. The following section details the formulation of the multiple-hyperplane (MH) classification, considers its generalization properties and presents preliminary experimental results.

Appearing in *Proceedings of the workshop Machine Learning Techniques for Processing Multimedia Content*, Bonn, Germany, 2005.

2. Multiple-hyperplane classification

2.1. Problem formulation

The standard 2-class optimal separating hyperplane problem setting can be extended trivially in order to accommodate more than one hyperplane:

$$\min_{\omega_1,\ldots,\omega_{N_H}} \quad ||\omega_1||^2 \tag{1}$$

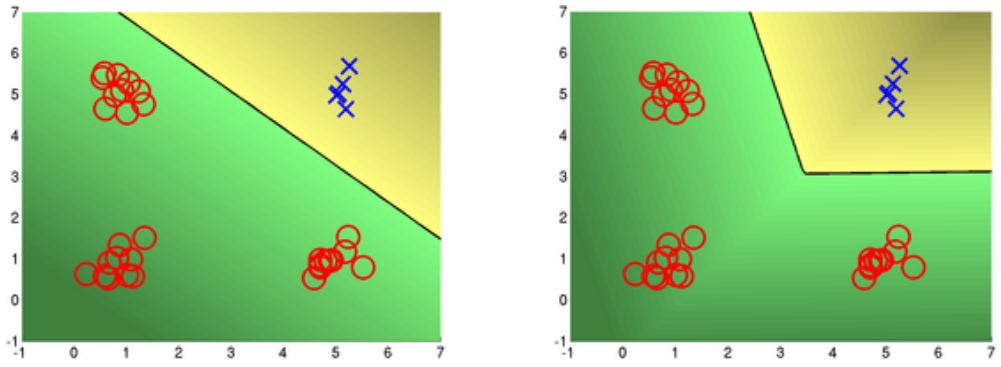
$$\text{subject to:} \quad y_i \min_{j=1\ldots N_H} (\omega_j^T x_i) \geqslant 1, \tag{2}$$

$$||\omega_1||^2 = \ldots = ||\omega_{N_H}||^2, \tag{3}$$

where $(x_i, y_i) \in \mathbb{R}^n \times \{\pm 1\}$ are data samples with their respective class labels, and N_H is the number of hyperplanes, each of which is defined by ω_j. Here, labels $+1$ and -1 correspond to under-represented and over-represented classes respectively. Additionally, we require that the sum of distances to compound border be less or equal to the sum of signed distances to the average hyperplane $\bar{\omega}$:

$$\sum_i y_i \min_{j=1\ldots N_H} (\omega_j^T x_i) \leqslant \sum_i y_i \bar{\omega}^T x_i, \tag{4}$$

where $\bar{\omega} = \frac{1}{N_H} \sum_{j=1\ldots N_H} \omega_j$. This condition ensures some degree of flatness of the compound border avoid-



(a) SVM, $\gamma = 1.14$ (b) MH, $\gamma = 1.55$

Figure 1. SVM vs. Multiple-hyperplane (MH) method on a toy problem in 2D: an additional hyperplane leads to a better separation margin γ (both methods use linear kernels).

ing overfitting. There is guaranteed to be at least one set of hyperplanes that meets this requirement. The actual role of this average signed distance constraint, however, will be clarified in greater detail in the following section.

A disadvantage of the proposed formulation is that the above optimization problem may be quite diffcult due to the use of non-differentiable min-function, which necessitates the use of auxiliary numerical strategies for attaining differentiability via smoothing of the loss function and avoiding unacceptable local minima via annealed penalty terms. Its advantage, on the other hand, is that (1-4) are expressed in terms of dot products, and thus are easily extended to nonlinear cases via kernel trick.

2.2. Generalization performance assessment

The following result, which we state without a detailed proof due to space limitations, establishes the generalization properties of the proposed technique.

Proposition 1. *Consider thresholding a class* $\mathbf{F}$ *of functions* $\min_{j=1\ldots N_H}(\omega_j^T t)$ *with unit weight vectors on an inner product space* $\mathcal{X}$ *and fix* $\gamma \in \Re^+$. *For any probability distribution* $\mathcal{D}$ *on* $\mathcal{X} \times \{-1, 1\}$ *with support in a ball of radius* R *around the origin, with probability* $1-\delta$ *over* l *random examples* S, *any hypothesis* $f \in \mathbf{F}$ *that has margin* $m_S(f) \geq \gamma$ *on* S *has error no more than*

$$\varepsilon(l, \mathbf{F}, \delta, \gamma) = \frac{2}{l}\left(\frac{64R^2}{\gamma^2}\log\frac{el\gamma}{4R}\log\frac{128lR^2}{\gamma^2} + \log\frac{4}{\delta}\right), \tag{5}$$

provided $l > 2/\varepsilon$ *and* $64R^2/\gamma^2 < l$.

Note that the error bound is absoultely the same as presented in (Cristianini & Shawe-Taylor, 2000) for a single hyperplane case. In order to clarify the intuition behind this result, we observe that the proof of a standard result on fat-shattering dimension, $fat_{\mathbf{F}}$, of an optimal hyperplane classifier (Schölkopf & Smola, 2002; Bartlett & Shawe-Taylor, 1999; Vapnik, 1982) is applicable in the multiple-hyperplane setting (1-4). That is, proceeding in a manner similar to the original proof and explicitly taking constraint (4) into account leads to an identical bound on $fat_{\mathbf{F}}$:

$$r^2\gamma^2 N_H \leqslant \left\|\sum_i^r y_i S^T x_i\right\|^2 \leqslant N_H r R^2$$
$$\Rightarrow \quad fat_{\mathbf{F}}(\gamma) \leqslant r \leqslant \left(\frac{R}{\gamma}\right)^2. \tag{6}$$

Then, result (5) naturally follows, once (6) is substituted into the theoretical result that establishes the link between the fat-shattering dimension and generalization error (Bartlett & Shawe-Taylor, 1999; Vapnik, 1982). In equation (6) above, r is the number of observations x_i, R is the radius of the smallest sphere containing all x_i, γ is the separation margin, and $S = \mathbf{1}^T \otimes I$ for a vector $\mathbf{1}$ of all ones of length N_H.

2.3. Preliminary experimental results

For our content-based multimedia retrieval experiments we chose ETHZ80 collection (Leibe & Schiele, 2003), containing 3280 high-resolution color images of objects from 8 different semantic classes. The visual information for each image was represented by 286-dimensional feature vector containing 166 global color histogram and 120 Gabor filter texture descriptors.

Table 1. Classification accuracy (in %%) per class for ETHZ80 image collection

Method	**apple**	**car**	**cow**	**cup**	**dog**	**horse**	**pear**	**tomato**
MH classifier	97.12	88.44	89.75	95.41	92.37	88.44	95.19	98.38
N_H	6	2	5	5	6	4	5	2
SVM classifier	96.16	88.06	84.59	95.94	83.59	88.09	92.16	97.66
margin ratio (MH/SVM)	1.37	1.10	1.11	1.02	1.77	1.76	1.22	1.01

For each semantic class the training data comprised 80 images with an imbalance ratio of 10/70, and an overall training vs. testing data ratio was hence 80/3200. For each class, we compared the classification accuracy of the 2-class SVM (Cristianini & Shawe-Taylor, 2000; Vapnik, 1998) with a Gaussian kernel tuned by cross-validation to that of the MH classifier using the same kernel parameters, but letting the number of hyperplanes vary. The outcome of these experiments demonstrated that in most cases the performance of the SVM classifier is improved by introducing extra separating hyperplanes, while the ratio of the class separation margins achieved by the two methods indicated where such improvement was most likely. The summary of results is shown in Table 1.

3. Conclusion

We have presented a large margin classification method that exploits the asymmetric problem setting by increasing the number of hyperplanes used in an optimal separating hyperplane classifier. The performance of the proposed technique has been assessed theoretically by establishing a bound on generalization error, and practically by evaluating its performance in a content-based image retrieval task, providing encouraging results. Further research is warranted in order to gain a better insight into the method's theoretical properties via Rademacher complexity bounds (Bartlett & Mendelson, 2001; Koltchinskii & Panchenko, 2002), and to investigate its performance in related multimedia processing applications.

References

Akbani, R., Kwek, S., & Japkowicz, N. (2004). Applying support vector machines to imbalanced datasets. *Proceedings of the 15th European Conference on Machine Learning (ECML'04)* (pp. 39–50).

Bartlett, P. L., & Mendelson, S. (2001). Rademacher and Gaussian complexities: Risk bounds and structural results. *Proceedings of the Fourteenth Annual Conference on Computational Learning Theory and Fifth European Conference on Computational Learning Theory* (pp. 224–240).

Bartlett, P. L., & Shawe-Taylor, J. (1999). Generalization performance of support vector machines and other pattern classifiers. In B. Schölkopf, C. J. C. Burges and A. J. Smola (Eds.), *Advances in kernel methods – support vector learning*, 43–54. MIT Press.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press.

Koltchinskii, V., & Panchenko, D. (2002). Empirical margin distributions and bounding the generalization error of combined classifiers. *Ann. Statist.*, *30*, 1–50.

Leibe, B., & Schiele, B. (2003). Analyzing appearance and contour based methods for object categorization. *International Conference on Computer Vision and Pattern Recognition (CVPR'03)* (pp. 409–415). Madison, Wisconsin.

Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. Cambridge, MA: MIT Press.

Vapnik, V. N. (1982). *Estimation of dependencies based on empirical data*. New York: Springer-Verlag.

Vapnik, V. N. (1998). *Statistical learning theory*. New-York: Wiley.

Veropoulos, K., Cristianini, N., & Campbell, C. (1999). Controlling the sensitivity of support vector machines. *Proceedings of the International Joint Conference on Artificial Intelligence, (IJCAI99)* (pp. 55–60). Stockholm, Sweden.

Zhou, X., Garg, A., & Huang, T. (2004). A discussion of nonlinear variants of biased discriminants for interactive image retrieval. *Proceedings of CIVR'04* (pp. 353–364). Dublin, Ireland.

AdaBoost learning of shape and color features for object recognition

Thang V. Pham VIETP@SCIENCE.UVA.NL
Arnold W. M. Smeulders SMEULDERS@SCIENCE.UVA.NL
Sanne Ruis
ISIS - University of Amsterdam,
Kruislaan 403, 1098 SJ Amsterdam, the Netherlands.

Abstract

We explore the problem of recognizing objects in images dominated by a broad background, aiming for a low false alarm rate. We propose a new discriminative model which consists of a large number of weighted object indicators. The model is learned efficiently and automatically from pictorial examples using the AdaBoost algorithm, exploiting both color and shape information. Unlike previous methods we compensate for the class priors in the training data by assigning unequal weights and updating them according to classification error. We perform experiments on several datasets including pedestrians, cars, and the COIL-100 color objects. The experimental results demonstrate that one algorithm with (almost) no parameter tuning can learn to recognize objects and perform as well as methods dedicated to each problem individually.

1. Introduction

Trainable object recognition is a challenging problem in computer vision. First of all, the appearance of the object makes the learning problem not trivial for a number of reasons. The training examples often lie in a high dimensional space. In many cases, the background is also present in the object examples. And to make the matter worse, even though the training data is often seriously under sampled from the underlying distribution, a reasonable number of training examples can be very large. In addition, generative object models are not known and in fact not meaningful for a generic object class. The second issue is the class imbalance problem. In a typical scene, the number of instances of the object class is much less than that of the background class. The training data, therefore, might not represent the class prior correctly. Finally, it is desirable that minimal supervision is needed in training. This includes manual annotation of training data, feature selection and parameter tuning.

We explore how shape and color information can be used and combined to generate discriminative object representations. In particular, we propose a new object model capturing discriminative aspects between object shape and color and background at various image locations in space, scale and orientation. We do not attempt to evaluate different ensemble learning algorithms. We choose the AdaBoost algorithm (Freund & Schapire, 1997) because of its performance in practice and its ease of implementation. We perform experiments on a number of datasets including pedestrians, cars, and the COIL-100 color objects. The experimental results demonstrate the genericity and efficiency of the new method.

The paper is organized as follows. In the next section we review related methods in literature. We then describe the AdaBoost algorithm in section 3. In section 4 we present the new discriminative object model and describe a solution to the class imbalance problem in model learning. Our experiments are presented in section 5. Section 6 concludes the paper.

2. Related Work

Sung and Poggio (Sung & Poggio, 1998) and Rowley et al. (Rowley et al., 1998) present early trainable systems in the face detection domain. The former assume a mixture of Gaussians for both object and background classes while the latter use a multilayer neural network. A number of methods follow with different learning algorithms. The major obstacle to a generic object detection system lies in their exploration of training

Appearing in *Proceedings of the workshop Machine Learning Techniques for Processing Multimedia Content*, Bonn, Germany, 2005.

data. The performance of appearance-based methods, where all pixel values are used in classification, is likely to degrade when background is embedded with object examples. In addition, the scalability of the learning techniques has to be examined because of the large size and/or high dimensionality of the training data for other object classes rather than human faces. Finally, the "bootstrap" method of collecting negative examples (Sung & Poggio, 1998) is not easy to automate for generic object classes, for example the setting of the accuracy of the system in each bootstrapping round.

Viola and Jones (Viola & Jones, 2004) present a fast object detection system by using a cascade of classifiers. This type of classifiers provides a viable approach to exploring negative examples. However, training an optimal classifier of this type is extremely difficult (Viola & Jones, 2004). Thus, a heuristic approach is adopted. As a result, the generalization performance of a cascade of classifiers is not clear.

The works in (Burl & Perona, 1996; Ioffe & Forsyth, 2001; Agarwal & Roth, 2002) belong to the class of detection-by-part methods. In this approach the object parts are first detected, then grouped to form objects according to an explicit spatial relationship among parts. This approach is intuitive. However, under the presented formulation only translation of parts is dealt with. The difficult problem of learning the object model is addressed in (Weber et al., 2000; Ioffe & Forsyth, 2001; Agarwal & Roth, 2002) where the specific object class is handled.

Detection by part can be seen from a different perspective. Mohan et al. (Mohan et al., 2001) model the pedestrian object class by six components. The support vector machine learning method (Vapnik, 1998) is used to train a detector for each component and to train a combined classifier. The system shows robust detection even when partial occlusion occurs, which is a clear advantage of this approach. Their result also shows that combination of classifiers outperforms a single classifier approach. The major drawback is that the model is constructed manually. This problem is in fact also present in their related work (Papageorgiou & Poggio, 2000) where a reduced subset of features is selected manually to improve the detection speed.

In summary, methods that assume a generative model are not suitable for generic detection system, while distribution free methods such as support vector machine (SVM) (Vapnik, 1998) or Sparse Network of Winnow (SNoW) (Yang et al., 2000) do not fully address the class imbalance problem in an automatic manner. In addition, appearance-based methods do not provide a viable solution to the problem where background is embedded with object examples. The detection by part approach deals with this problem elegantly. However, the problem of learning a generic object model remains unsolved. Furthermore, current methods consider object parts at one scale and orientation only, and hence important discriminative features might not be used.

3. AdaBoost learning

Let us consider a standard two class classification problem. Let there be a training set $\{(x_i, y_i)\}$ drawn from some fixed but unknown distribution $P(x, y)$ on $X \times Y$, where X is the space of the data variable x and $Y = \{-1, 1\}$ is the set of the class label y. In our context, -1 denotes the background class and 1 denotes the object class. The task is to predict the label y given x.

Among the various learning techniques, ensemble learning methods (Freund & Schapire, 1997; Breiman, 1998) are suited for our problem because they are efficient and robust with respect to training data while making no assumption about the underlying distribution. The fact that they work directly in the distribution space of the input data allows us to deal with the class imbalance problem in a simple manner. They are flexible in that prior knowledge can be incorporated via the class of base classifiers. This allows us to design a discriminative model combining both color and shape information.

In this paper we are interested in a class of ensemble methods which finds a sparse linear combination of base classifiers (Freund & Schapire, 1997). Specifically, suppose that there are a set of classifiers (weak hypotheses) $\mathcal{H} = \{h_t : X \rightarrow Y\}$ and a learning algorithm (base learner) which returns a hypothesis $h_t \in \mathcal{H}$ for any distribution over the inputs. The number of classifiers in $\mathcal{H}$ could be infinite. A classifier ensemble is constructed by iteratively calling the base learner with an appropriate distribution, depending on the empirical performance of the hypotheses learned in the previous steps.

The AdaBoost algorithm (Freund & Schapire, 1997) is a powerful ensemble learning method. Empirical studies in (Breiman, 1998) show that the performance of the AdaBoost algorithm is similar or slightly better than related ensemble methods in terms of generalization. We choose the AdaBoost algorithm because of the ease of implementation. A summary of the algorithm is as follows.

THE ADABOOST ALGORITHM

Input: N examples $\{(x_i, y_i)\}$ and an initial distribution represented by a set of weights $D_1(i)$ over the examples.

Do for $t = 1, \ldots, T$

1. Learn a hypothesis $h_t \in \mathcal{H}$ from the training examples with distribution D_t.
2. calculate the empirical error of h_t

$$\epsilon_t = \Pr_{i \sim D_t}[h_t(x_i) \neq y_i] \tag{1}$$

3. set

$$\alpha_t = \frac{1}{2} \ln \left(\frac{(1 - \epsilon_t)}{\epsilon_t} \right)$$

4. update

$$D_{t+1}(i) = \frac{D_t(i) e^{(-\alpha_t y_i h_t(x_i))}}{Q_t}$$

where Q_t is a normalization factor.

Output: The final classifier

$$f_{\mathcal{H}}(x) = \text{sign}\left(g_{\mathcal{H}}(x)\right) \tag{2}$$

where

$$g_{\mathcal{H}}(x) = \sum_{t=1}^{T} \alpha_t h_t(x) \tag{3}$$

$g_{\mathcal{H}}(x)$ might be used to indicate the confidence of the classification.

4. Discriminative Object Model

A set of hypotheses $\{h_t\}$ together with their weights $\{\alpha_t\}$ and the discriminant function eq. (2, 3) serve as an object model.

In this section we present a class of weak hypotheses $\mathcal{H}$, which we call object indicators, and a base learning algorithm which returns an object indicator for each distribution over the training examples. Finally, we describe a simple way to deal with the class imbalance problem.

4.1. Object Shape Indicators

Each object indicator serves as a cue suggesting the presence of an object instance. We use the intensity changes at different image locations in space, scale and orientation. The indicator is local at one scale, but global at a finer scale.

First of all, we transform the input image into a new representation. An image pyramid is constructed. Each level of the pyramid is smoothed with a Gaussian kernel, then convoluted with a Gaussian derivative filter in two orthogonal directions. At each spatial location, the strength and direction of the response are computed from the two convolutions. The strength is then thresholded. If it is above a threshold value, the response direction is discretized.

We consider a class of local object indicators of the form $\mathcal{H} = \{h^{(\boldsymbol{l},d,s|\boldsymbol{r_s})} : X \to Y\}$ where $\boldsymbol{l}$ denotes a spatial image location, d a response direction, s a level of the pyramid (scale) and $\boldsymbol{r_s}$ the size of a neighborhood of $\boldsymbol{l}$ at level s. Note that both $\boldsymbol{l}$ and $\boldsymbol{r_s}$ are vectors. An object indicator $h^{(\boldsymbol{l_0},d_0,s_0|\boldsymbol{r_{s_0}})}$ is defined to classify an input pattern $x \in X$ as object ($y = 1$) if there is an intensity change in direction d_0 in the $\boldsymbol{r_{s_0}}$-neighborhood of $\boldsymbol{l_0}$ at level s_0 of the pyramid, and as background ($y = -1$) otherwise. Figure 1 shows an example of an object indicator at level s of the pyramid.

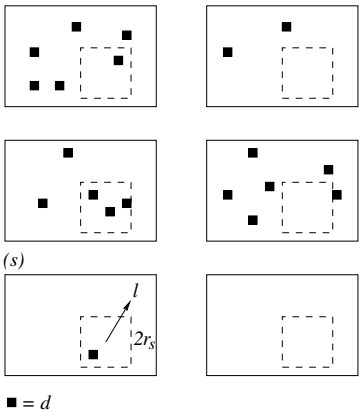


Figure 1. An example of an object indicator $h^{(\boldsymbol{l},d,s|\boldsymbol{r_s})}$ for six different patterns at one scale shown with a given neighborhood $\boldsymbol{r_s}$. The black squares indicate the locations where an intensity change in direction d occurs. This indicator classifies the patterns on the left as object instances ($y = 1$) and ones on the right as background ($y = -1$).

4.2. Object Color Indicators

We take a simple approach for color using color histogram. Currently, in our experiment, we use the normalized rg-histogram. An indicator is constructed from each bin in the histogram. Given a set of training samples with a weight distribution for each bin b_j a threshold is calculated. This threshold separates the positive and the negative samples as well as possible. This is done using the following equation similar to

(Viola & Jones, 2004)

$$h_j(x) = \begin{cases} 1 & \text{if } p_j b_j(x) > p_j t_j \\ -1 & \text{otherwise.} \end{cases} \quad (4)$$

where p_j is a parity indicating the direction of the inequality sign, and t_j a threshold value.

4.3. Indicator Learning in Distribution Space

The task of the base learner is to find an indicator $h \in \mathcal{H}$ for each set of weights representing a distribution over the training examples. Our approach is to perform an exhaustive search over $\mathcal{H}$ to find the one with the lowest empirical error. The computational cost can be reduced by sampling from the training data according to their weights. Then the unweighted sampled data can be used for training. In our experiments the exhaustive search approach is feasible.

4.4. Model Learning in Class Imbalance

The indicator space $\mathcal{H}$ and the base learning algorithm can be used in the AdaBoost algorithm to learn the object model as in section 3. The problem arises, however, when training examples are not drawn from the joint distribution $P(x, y)$, but obtained separately. This situation is rather common in object detection. Typically, a number of instances of the object class are prepared. They should be representative for the object class, and ideally are drawn randomly from $P(x|y = 1)$. Similarly, the set of background instances is drawn from $P(x|y = -1)$. So, the training set is obtained from the two class conditional distributions.

In the context of object detection one is interested in two types of misclassification, namely the rate of missing true objects ϵ_I and the rate of false alarms ϵ_{II}. Let ϵ denote the generalization error. We have

$$\epsilon = \epsilon_I \lambda + \epsilon_{II}(1 - \lambda) \quad (5)$$

where $\lambda = P(y = 1)$ and $(1 - \lambda) = P(y = -1)$ are the class priors. Thus, a learning algorithm aiming at lowering the generalization error ϵ also drives down the false alarm rate ϵ_{II} when proper class priors are provided. A small value of λ, or equivalently a large value of $(1 - \lambda)$, leads to a very low false positive rate ϵ_{II}.

A simple way to correct the class prior in the training set is to set the weights $D_1(i)$ of the examples equal in each class and sum up λ and $(1 - \lambda)$, respectively. A dataset sampled from the training set according to the weights $D_1(i)$ will reflect the class priors properly.

Note that λ is an input parameter, reflecting the user prior belief on the class ratio. It is similar to the final decision threshold used in many systems, which is usually set to $P(y = 1)/P(y = -1)$. Unlike these systems, the parameter λ is used in training in our approach. As a result, generating a ROC curve is computationally expensive since we need to re-train for each performance target. However, the advantage is that better performance is achieved by using a proper training set.

In short, we achieve a very low false positive rate by adjusting the initial weights of the training examples to reflect the class priors. Significantly, unlike the bootstrapping or cascade of classifiers approaches, this method maintains generalization properties of the learning algorithm. Furthermore, there is only one parameter to be specified and hence the learning step is fully automatic.

5. Experiment

This section presents our experiments on a number of datasets. First, we summarize the parameters of the system. We then give a brief description of each dataset and the experimental results. Finally we show the object models learned in our experiments.

5.1. System Parameters

The scale is enlarged 20% each iteration over scale. The size of the neighborhood at each level of the pyramid is proportional to the size of that level. A small value is expected at the top level of the pyramid, for example 0.5. Although setting the threshold is problematic due to the differences in image contrast, it appears that camera and balancing of contrast function similarly. In our experiments, a value of 5 is used. A small deviation from this value does not effect the performance. Finally, the performance is not sensitive to small change in the number of discrete orientations. Its typical values are 4, 8 and 16.

There are two parameters in the learning phase. The first one is the class prior factor λ. It is application dependent. The second parameter is the number of iterations T. It could be fixed beforehand. But normally, it is set according to the performance of the classifier ensemble.

5.2. Datasets and Experimental Results

Pedestrian - DaimlerChrysler

This dataset is a subset of the one used in (Gavrila & Giebel, 2001). The full dataset is not available. It consists of 1500 pedestrian shapes (including mirroring) and 5000 background images of size 100×100. An example of each class is shown in figure 2. We use

500 examples in each class for training, leaving a test set of 1000 pedestrian and 4500 background images. There is no mirrored image of the training set in the test set. Exactly the same learning algorithm is used.

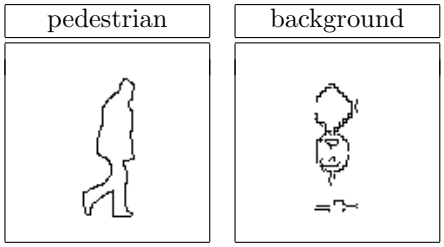


Figure 2. Examples of the DaimlerChrysler dataset.

Figure 3 shows the performance of the system. The system with 50 indicators classifies correctly 896 out of 1000 patterns in the object class with only 2 false alarms (a false alarm rate of 4.4×10^{-4}). This compares favorably to the result reported in (Gavrila & Giebel, 2001), where on a larger dataset the matching method using the chamfer distance achieves this detection rate at a false alarm rate of approximately 2.5×10^{-3}.

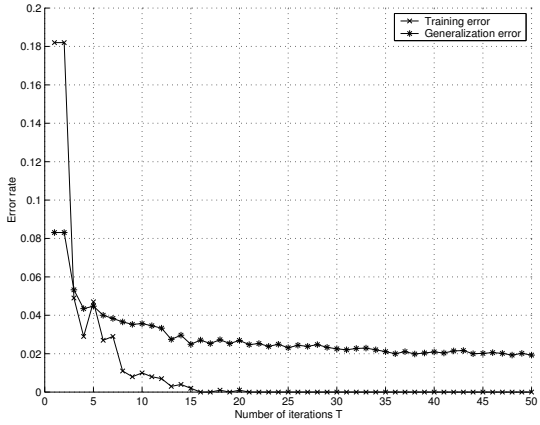


Figure 3. Performance on the DaimlerChrysler dataset.

Car - UIUC

The same algorithm is evaluated on the dataset used in (Agarwal & Roth, 2002). The dataset consists of 1050 training examples with 550 car images and 500 background images of size 100×40 (see figure 4). In addition, a test set of 170 images with 200 cars is available. The cars in the test images are of approximately the same scale as that of the training examples.

Table 1 shows the results of our system in comparison with (Agarwal & Roth, 2002) using the evaluation software generously provided by the authors of (Agarwal & Roth, 2002). The same number of windows is tested in this experiment as in (Agarwal & Roth, 2002). We also use the same neighborhood size for merging overlapping hypotheses, where eq. (3) is used as the confidence measure. The result demonstrates that our algorithm performs equally well in comparison with the system dedicated to this dataset.

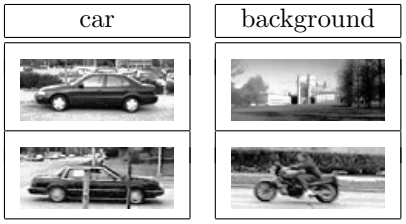


Figure 4. Examples of the UIUC car dataset.

	our system			UIUC system		
λ	TDs	DR%	FAs	TDs	DR%	FAs
10^{-5}	182	91.0	90	181	90.5	98
10^{-6}	178	89.0	64	178	89.0	92
10^{-7}	173	86.5	43	171	85.5	76
10^{-8}	164	82.0	24	162	81.0	48
10^{-9}	153	76.5	13	154	77.0	36
10^{-10}	142	71.0	5	140	70.0	29

Table 1. Results on the UIUC test set. TDs is the number of instances detected out of 200. DR is the detection rate, and FAs is the number of false alarms.

Figure 5 shows two examples of the detection results of our algorithm using the evaluation software in (Agarwal & Roth, 2002). Both cars are detected in the image on the left. There are one correct detection and one false alarm for the image on the right.

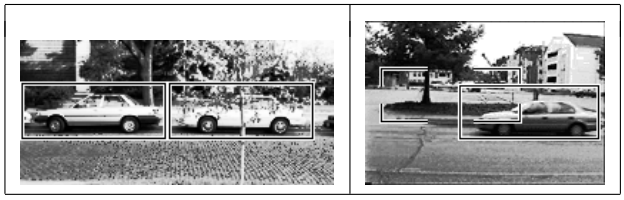


Figure 5. Examples of the detection result.

COIL-100

The experiments are performed on the COIL-100 dataset (Nene et al., 1996). This dataset contains 100 objects; for each object 72 views were taken 5 degree

apart. We performed experiment with different number of training views. Specifically, v views are used for training for each object and $(72 - v)$ views are used for testing. To train one object model, the training examples of other objects serve as the negative set. The confidence value in eq. (3) is used for multiclass classification.

We used histograms with 64 bins, 8 in each r and g directions.

Views per Object	36	18	8	4
SNoW	95.81	92.31	85.13	81.46
Linear SVM	96.03	91.03	84.80	78.50
Nearest neighbor	98.50	87.54	79.52	74.63
Shape only	95.61	91.44	78.34	58.96
Color only	97.89	97.17	92.41	77.72
Shape & Color	99.89	99.70	97.88	80.85

Table 2. Correct recognition rate (%) on the COIL-100 dataset (Nene et al., 1996).

Table 2 compares the recognition rate of our method and other learning algorithms. For the setting and result of previous experiments for SNoW, support vector machine, and nearest neighbor, the reader is referred to (Roth et al., 2002). As expected, the combination of shape and color outperforms shape and color feature alone. The combined shape and color model also compares favorably to all other algorithms in case of 36, 18 and 8 training views. In particular, in case of 36 training views, it gives a remarkable accuracy: only 4 test views were misclassified out of 3600 test views.

6. Discussion and Conclusion

We have proposed a new object model consisting of a large number of object indicators which capture discriminative aspects between the object and the background class at different image locations in space, scale, orientation, and also in color. The model is learned efficiently and automatically using the AdaBoost algorithm where the class imbalance problem is handled in a simple manner by adjusting the distribution over the training examples.

The experimental results on several datasets demonstrate the genericity of the method. Also, the various sizes of the datasets show the efficiency of the algorithm. The system achieves a very low false positive rate. In other words, it is able to correct the class prior in the training set. In terms of generalization, the system performs equally well in comparison to state of the art methods in object recognition, for examples (Gavrila & Giebel, 2001; Agarwal & Roth, 2002). We also demonstrated the ease of combining color and shape information with classifier ensemble, thereby giving a better performance than using individual modality.

Acknowledgments

This research is supported by the MUSCLE project and MultimediaN.

References

Agarwal, S., & Roth, D. (2002). Learning a sparse representation for object detection. *Proc. of ECCV* (pp. 113–130).

Breiman, L. (1998). Arcing classifiers (with discussion). *Annals of Statistics*, *26*, 801–849.

Burl, M., & Perona, P. (1996). Recognition of planar object classes. *Proc. of CVPR* (pp. 223–230).

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, *55*, 119–139.

Gavrila, D. M., & Giebel, J. (2001). Virtual sample generation for template-based shape matching. *Proc. of CVPR* (pp. 676–681).

Ioffe, S., & Forsyth, D. (2001). Mixtures of trees for object recognition. *Proc. of CVPR*.

Mohan, A., Papageorgiou, C., & Poggio, T. (2001). Example-based object detection in images by components. *IEEE PAMI*, *23*, 349–361.

Nene, S. A., Nayar, S. K., & Murase, H. (1996). *Columbia object image library (COIL-100)* (Technical Report Technical Report CUCS-006-96). Columbia University.

Papageorgiou, C., & Poggio, T. (2000). A trainable system for object detection. *IJCV*, *38*, 15–33.

Roth, D., Yang, M.-H., & Ahuja, N. (2002). Learning to recognize three-dimensional objects. *Neural Computation*, *14*, 1071–1103.

Rowley, H. A., Baluja, S., & Kanade, T. (1998). Neural network-based face detection. *IEEE PAMI*, *20*, 23–38.

Sung, K. K., & Poggio, T. (1998). Example-based learning for view-based human face detection. *IEEE PAMI*, *20*, 39–51.

Vapnik, V. N. (1998). *Statistical learning theory.* John Wiley & Sons, Inc.

Viola, P. A., & Jones, M. J. (2004). Robust real-time face detection. *IJCV*, *57*, 137–154.

Weber, M., Welling, M., & Perona, P. (2000). Unsupervised learning of models for recognition. *Proc. of ECCV* (pp. 18–32).

Yang, M.-H., Roth, D., & Ahuja, N. (2000). A snow-based face detector. *NIPS 12* (pp. 855–861). MIT Press.

A Bayesian Method for Automatic Landmark Detection in Segmented Images

Katarina Domijan DOMIJANK@TCD.IE
Department of Statistics, University of Dublin, Trinity College, Dublin 2, Ireland

Simon Wilson SWILSON@TCD.IE
Department of Statistics, University of Dublin, Trinity College, Dublin 2, Ireland

Abstract

The identification of landmark points of a figure in an image plays an important role in many statistical shape analysis techniques. In certain contexts, manual landmark detection is an impractical task and an automated procedure has to be employed instead. Standard corner detectors can be used for this purpose, but this approach is not always suitable, as the set of landmark points best representing the figure is not necessarily limited to corners. We present a Bayesian approach for automatic landmark detection, where a set of N landmark vertices is fitted to the edge of a segmented region of an image. We propose a likelihood function for the observed segmented region given the vertices and then use a Metropolis sampler to sample landmark vertices given the observed region. Careful consideration has to be given to the selection of a prior for the distribution of the landmarks.

1. Introduction

The shape of an object in a two dimensional image is often characterized by a set of N labelled points and hence is represented by an $N \times 2$ matrix. This type of shape representation scheme can be extended to R dimensional surfaces and it satisfies the requirements of invariance to translation, scale and rotation. It therefore is a basis for many shape analysis methods. Such data arises in many applications and the corresponding labelled points are commonly called landmarks. In certain applications, for example in biological homology, landmarks are assumed to be uniquely defined locations that are identifiable across a particular class of objects or individuals. In general, it is assumed that a set of landmarks is found in at least two objects and the interest is focused on their relative positions. This abstraction allows shape theory to stand apart from issues of interpretation (Goodall, 1991).

In object recognition context, one does not a priori know the class of objects that the region of interest belongs to. Therefore, this kind of definition is not applicable. For this purpose, we define landmarks to be be a set of coordinate points that best describe a given region. The distinction between landmarks of an object and salient points of an image is that the purpose of salient points is not to summarize the shape contour of an object, but rather to represent a subset of image pixels where the image information is supposed to be most important (Sebe & Lew, 2003).

Manual landmark detection is too time-consuming in content based image retrieval applications where one might be dealing with large databases of images. Arguably it is also too subjective (Brett & Taylor, 2000). In segmented images where a region contour is clearly defined it is possible to use corner detectors such as Harris (Harris & Stephens, 1988), as well as a number of other algorithms. For the purposes of image retrieval, it is interesting to obtain information on the uncertainty of the shape retrieved, which is why a Bayesian approach is useful.

In this paper, presented is a Bayesian method for automatic detection of landmarks in pre-segmented images. The idea is to fit a set of N landmark vertices to the edge of a segmented region of interest, with the aim of describing the shape of that region well. The edge of the region is taken to be the object contour. There is a restriction for the segmented region to be

Appearing in *Proceedings of the workshop Machine Learning Techniques for Processing Multimedia Content*, Bonn, Germany, 2005.

solid, i.e. without holes. In theory, the particular segmented region would represent one object of interest in that image.

The Bayesian framework requires a likelihood function to be proposed for the observed segmented region given the landmark vertices and then a Metropolis sampler is used to sample landmark vertices given the observed region. Hence we obtain a distribution for the set of landmarks given the segmented region from which we can draw inferences on the landmarks set. In the following section of the paper this model is described in more detail. Subsequently, the method was applied to an artificial test example and a pre-segmented image of a painting from the Bridgman Art Library, London.

Two main Bayesian approaches to high level imaging, which involves working with components of an image in such tasks as object recognition are based on pattern theory (Grenander & Miller, 1994) and marked point processes (Baddeley & van Lieshout, 1993); for a recent contribution see (Hurn, 1998). The first approach uses a deformable template to represent the outline of a typical object and the natural variability is often represented by a probability measure on the parameters affecting the deformations. Kent et al. (2000) further consider some statistical aspects of this approach including maximum likelihood based. In the second approach, the images are characterized by processes of simple geometrical figures, each specified by a location and a mark containing information such as the shape and size of the figure. Rue and Hurn (1999) combine these two approaches by imbedding the template models into a marked point process framework. Other work has been done in estimating object boundaries in an image, usually with some prior knowledge of the object shape. However, these approaches differ from the one discussed in this paper in that they seek to obtain a contour of an object, as opposed to selecting a set of points that best represent an already estimated contour shape obtained from a segmentation of the image.

2. The Bayesian Model

The Bayesian approach to the problem of selecting a set of landmark points to best represent the shape of a segmented region in an image could be described as the following: the prior distribution for the scene of interest X, π(x), is combined with the likelihood of the data Y arising from a particular scene X, $\pi(y|x)$. In this particular case, X is a set of ordered N landmark points, where each point is specified by a two coordinate location vector in the image matrix. The data Y is a matrix of pixels in the segmented image, indexed as either belonging to the region of interest or not. Inferences for X are made using the posterior distribution

$$\pi(x|y) \propto \pi(y|x)\pi(x).$$

Hence π(x) is the prior distribution for the locations of landmarks and π(y|x) is the likelihood of the observed shape arising given the landmark points' locations. The rest of this section describes the model choices for π(y|x) and π(x).

2.1. Prior Distribution for X

The set of ordered landmark vertices forms a N-sided landmark polygon. Note that in this model, the number of vertices N is a constant which needs to be set by the user.

To model the fact that the landmark polygon edges are not permitted to cross over, one can specify the prior with the indicator function π(x) $\propto$ I [edges crossing].

The prior distribution does not place a restriction on the points to be on the edge of the segmented region. Also the points need not be equally spaced, as this restriction may not always result in landmarks best describing the segmented region.

2.2. Likelihood

One possible data model is an increasing function of the distance of the pixels from edge of the landmark polygon. So the data model assumed is

$$\begin{aligned}\pi(y|x,\alpha) &= \prod_{pixels(s,t)\in S} \pi(y_{st} \in S|x,\alpha) \times \\ &\quad \times \prod_{pixels(s,t)\notin S} \pi(y_{st} \notin S|x,\alpha)\end{aligned}$$

where

$$\pi(y_{st} \in S|x,\alpha) = \begin{cases} \frac{1}{1+\exp(-\alpha\frac{d}{D})} & \text{if } y_{st} \in \text{L} \\ 1 - \frac{1}{1+\exp(-\alpha\frac{d}{D})} & \text{if } y_{st} \notin \text{L} \end{cases}$$

and

$$\pi(y_{st} \notin S|x,\alpha) = 1 - \pi(y_{st} \in S|x,\alpha).$$

S is the region of interest in the image, L is the region bounded by the landmark polygon, D is the largest minimum distance between the pixel and each edge in the landmark polygon and d is the smallest minimum distance. The likelihood term contains the unknown parameter α for which a uniform prior between 0 and a large upper bound is used.

Note that this simulation of the likelihood simply models the property that pixels from the polygon edge are

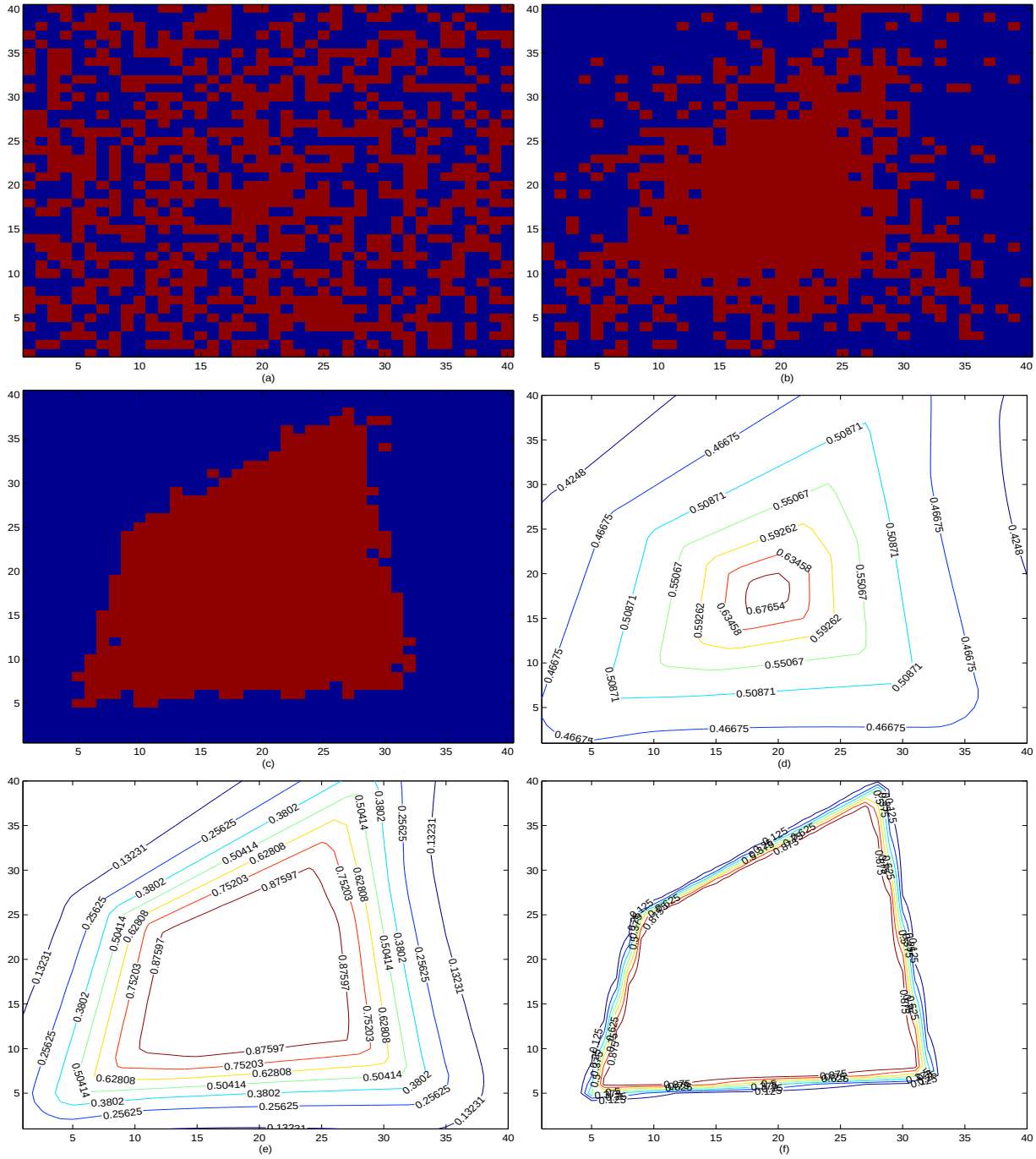


Figure 1. (a) Observed segmented region simulated for a set of landmarks with α=1, (b) α=10 and (c) α=80. (d) observed likelihood function contours for α=1, (e) α=10 and (f) α=80.

less likely to be classifies as being inside the shape. More complex likelihoods do not appear necessary.

Figures 1(a) to (c) show the observed segmented region which was simulated for a given (artificial) set of landmarks with different parameters (α=1, α=10, α=80). Figures 1(d) to (f) show the observed likelihood function contours. Hence the full posterior distribution is

$$\pi(x, \alpha|y) \propto \pi(y|x, \alpha)\pi(x)\pi(\alpha).$$

2.3. Inferences

The Metropolis algorithm (Metropolis et al., 1953) was used to obtain an iterative sequence of {x,α} that con-

verges in distribution to $\pi(x, \alpha|y)$. The approach used was to update x and α one at a time while the other one is held fixed. The conditional distributions for both variables can be derived from the posterior distribution, the distribution of primary concern being $\pi(x|y)$:

$$\pi(x|y, \alpha) \propto \pi(y|x, \alpha)\pi(x).$$

The candidate generating density for x was set to be multivariate normal, where at each iteration of the algorithm the location of only one vertex at a time was perturbed. The vertex to be perturbed was randomly chosen.

The Metropolis algorithm requires initial values to be provided for all the variables. For parameter α, a value greater than zero was randomly chosen. From a segmented image, a starting set of landmark points can be obtained by randomly selecting their locations in the image matrix, or by first using an edge detector to obtain the edge points of the shape of interest and then randomly sampling from the edge point locations to obtain a set of N landmark points. The randomly selected initial landmark points can be reordered by an algorithm such as the nearest neighbour.

3. Results

3.1. Artificial Test Example

An artificial image was created to illustrate the sampling behaviour of the model. The shape Of interest is a simple rectangular region. An initial set of landmarks (with N=4) was obtained by randomly selecting their locations in the image matrix.

A sequence of realisations from the π(x|y) is obtained once the convergence of the algorithm appears to have been reached. In order to assess the convergence four separate simulations were run with overdispersed starting points. Figure 2 shows the sequences for all 6000 iterations for the four simulation runs.

Figure 3(a) shows four starting landmark sets and the object of interest. Figure 3(b) shows the estimates (sample means) from the posteriors of the four landmark sets. Note that the first half of the iterations of the simulation runs was discarded for the purpose of making inferences from the posterior.

3.2. Bridgman Art Library Painting

One segmented region was chosen in a pre-segmented image from the Bridgman art library. The Prewitt edge detector (Prewitt & Mendelsohn, 1966) was used to identify the edge points of the region from which 15 points were randomly selected as the starting set of

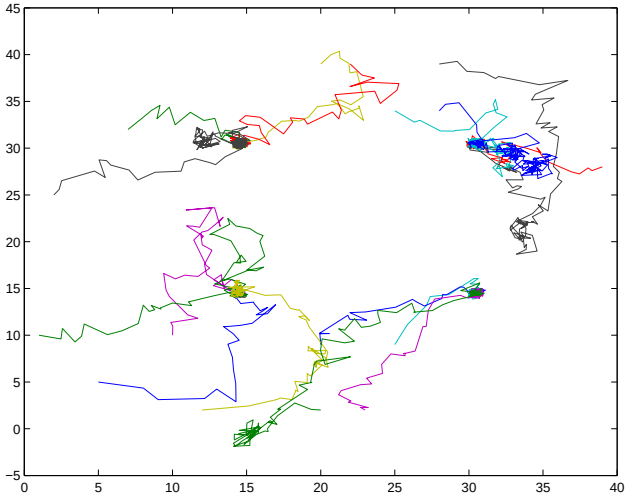


Figure 2. Four independent sequences of the simulations with different starting points. All 6000 iterations are plotted for each sequence and each landmark vertex. The starting points are indicated by crosses.

landmarks. Figure 4(a) shows the starting landmark set superimposed on the region and figure 4(b) shows the estimate of the landmark sets from the posterior distribution. Whereas the Metropolis algorithm seems to converge for the artificial test example, there are still some convergence and mixing problems with the more complicated shape.

4. Discussion

In this study, the problem of automatically generating a set of landmark points to describe the shape of a region of interest in segmented images has been attempted by using a Bayesian framework. The advantage of the Bayesian approach is that it provides information about the uncertainty of the shape, i.e. the uncertainty of how good the landmarks chosen are at summarizing the region of interest. This is particularly useful in content based image retrieval applications, which is the aim of the future research on this topic. This automatic landmark detection method will be implemented to a content based image retrieval application, where given a large database of segmented images, the shapes of segmented regions in different images are compared using Procrustes analysis.

5. Acknowledgements

Some of the images used in the development of this method are courtesy of the Bridgman Art Library London. This work is funded through the European Union

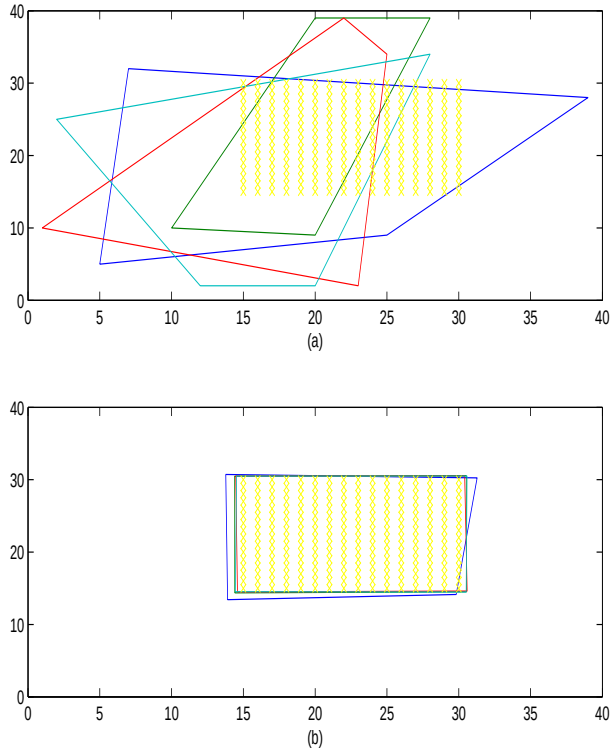


Figure 3. (a) Four starting landmark sets and the shape of interest. (b) Sample mean estimates of the four landmark sets.

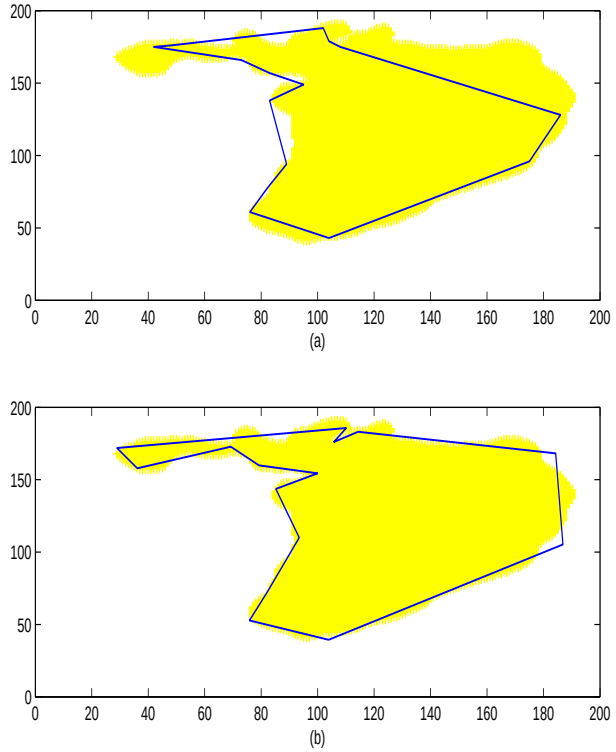


Figure 4. (a) Starting landmark set superimposed on the region. (b) Estimate of the landmark sets from the posterior distribution.

Network of Excellence MUSCLE.

References

Baddeley, A., & van Lieshout, M. (1993). Stochastic geometry models in high-level vision. In K. Mardia (Ed.), *Statistics and images*, 233–258. Abingdon: Carfax.

Brett, D., & Taylor, C. J. (2000). A method of automated landmark generation for automated 3d pdm construction. *Image and Vision Computing*, *18*, 739–748.

Goodall, C. R. (1991). Procrustes methods and the statistical analysis of shape (with discussion). *J. Royal Statistical Soc. B*, *53*, 285–340.

Grenander, U., & Miller, M. I. (1994). Representations of knowledge in complex systems. *J. Roy. Statist. Soc. B*, *56*, 549–603.

Harris, C., & Stephens, M. (1988). A combined corner and edge detector. *Proceedings of the 4th Alvey Vision Conference*, 147–151.

Hurn, M. (1998). Confocal flourescence microscopy of leaf cells: an application of bayesian image analysis. *Appl. Statist.*, *47*, 361–377.

Kent, J. T., Dryden, I. L., & Anderson, C. R. (2000). Using circulant symmetry to model featureless objects. *Biometrika*, *87*, 527–544.

Metropolis, N., Rosenbluth, N., Teller, M., & Teller, E. (1953). Equations of state calculations by fast computing machines. *J. Chemical Physics*, *21*, 1087–1092.

Prewitt, J. M. S., & Mendelsohn, M. L. (1966). The analysis of cell images. *Ann. N. Y. Acad. Sci.*, *128*, 1035–1053.

Rue, H., & Hurn, M. (1999). Bayesian object identification. *Biometrika*, *86*, 649–660.

Sebe, N., & Lew, M. (2003). Comparing salient point detectors. *Pattern Recognition Letters*, *24*, 89–96.