

TO APPEAR IN  
PATTERN  
RECOGNITION  
LETTERS

# Recognizing Image “Style” and Activities in Video Using Local Features and Naive Bayes \*

Daniel Keren

Department of Computer Science

University of Haifa

Haifa 31905, Israel

dkeren@cs.haifa.ac.il

## Abstract

The goal of this paper is to offer a framework for classification of images and video according to their “type”, or “style” – a problem which is hard to define, but easy to illustrate; for example, identifying an artist by the style of his/her painting, or determining the activity in a video sequence. The paper offers a simple classification paradigm based on local properties of spatial or spatio-temporal blocks. The learning and classification are based on the naive Bayes classifier. A few experimental results are presented.

**Keywords:** Image style, texture, naive Bayes, activity detection.

## 1 Introduction

One of the visual tasks which humans perform well may be described as “recognition by style”. For example, a person can quite successfully determine the identity

---

\*This research was supported by **The Israel Science Foundation** (grant No. 591/00-10.5)

of an artist given a hitherto not seen painting, if he/she is familiar with other paintings made by the artist, and if two artists with very similar styles are not present (still, in that case, it is possible to recognize the school of the painting – cubist, expressionist etc.).

The recognition of “style” does not use gray level or color similarity, nor high-level features (such as faces, eyes etc.), which excludes using many methods that are successful for other computer vision problems. Another interesting difficulty in the “style detection” problem is the construction of a training set, since, for example, every Dali painting is not “pure Dali”, and it will have some areas in it which appear as if they were painted by, say, Van-Gogh. Hence, the training sets of the positive and negative examples respectively will contain some negative and positive elements. The non-linear nature of the suggested method overcomes this difficulty, and also allows it to handle the case in which different styles are mixed (averaged).

This paper offers a simple, fast, and very easy to implement algorithm. It chooses local features which are based on the DCT transform coefficients, and then classifies the image/video blocks using the *naive Bayes* classifier. It may be viewed as a test as to how “very local” methods can identify style; in that regard, it is the opposite extreme of histogram-based methods such as the one described in [23]. Such local methods are important for cases in which a few “styles” co-exist closely together in an image or video sequence.

## 2 Previous Work

There is a large body of work related to the topic of this paper (which usually falls under the category of texture-based classification). The length of this paper allows to mention only some references to recent work. In [2], a mixture model was fitted to the output of a filter bank to classify shading and reflectance variations. [19] applies boosting to choose highly selective features for classification.

A sophisticated non-parametric multi scale texture analysis was presented in [6]. An application of global coefficient statistics to noise removal was offered in [17]. Images were classified by the rate of decay of their Fourier spectrum in [21]. More recent work is presented in [3, 10, 18, 1, 9, 22].

### 3 The Naive Bayes Classifier

The *naive Bayes* classifier is very popular in the data retrieval community, especially in text categorization applications [12, 7]. A short survey of the method's implementation follows.

1. A training set is given, which consists of a set of examples from the categories  $\{C_1, C_2, \dots, C_m\}$ . Denote the number of  $C_i$  examples as  $n_i$ , and the total number as  $\sum n_i = n$ . The probability of the  $i$ -th category is defined by  $P(C_i) = n_i/n$ . We shall refer to the examples as *texts*, although they do not necessarily have to be textual.
2. Define a set of possible *features*. In textual applications, these are usually words, classes of words which have a similar meaning, or "word stems". A feature may or may not appear in a text. For every feature  $f_i$  and category  $C_j$ , define  $P(f_i/C_j)$  as the ratio of  $C_j$ 's members which contain  $f_i$ , and  $P(f_i)$  as the ratio of all members of all categories which contain  $f_i$ . The important notion of *mutual information* between a feature  $f_i$  and category  $C_j$  is defined as

$$MI(f_i, C_j) = P(f_i/C_j) \log \left( \frac{P(f_i/C_j)}{P(f_i)} \right) \quad (1)$$

The mutual information has an attractive intuitive meaning; for it to be large, the frequency of  $f_i$  in  $C_j$  has to be large in absolute terms, and it also has to be large relative to  $f_i$ 's frequency in all the categories (its average frequency).

3. For every category, choose a few features which have the largest mutual information with respect to it. The union of these sets over all categories is called the *feature set*.
4. Given a new text  $T$ , extract all the features which it contains – call them  $\{f_{i_1}, f_{i_2} \dots f_{i_k}\}$  – and estimate for every category  $C_j$  the probability that  $T$  belongs to it, by

$$\begin{aligned}
P(C_j/T) &= \frac{P(C_j)P(T/C_j)}{P(T)} \approx \\
&\frac{P(C_j)P(\{f_{i_1}, f_{i_2} \dots f_{i_k}\}/C_j)}{P(\{f_{i_1}, f_{i_2} \dots f_{i_k}\})} \approx \\
&\frac{P(C_j) \prod_{l=1}^k P(f_{i_l}/C_j)}{\prod_{l=1}^k P(f_{i_l})} \tag{2}
\end{aligned}$$

The first equality is just Bayes’ law. The first approximation means that, when classifying  $T$ , we only consider the features it contains. The second approximation assumes that the presence of features is independent (this is where the “naive” in “naive Bayes” comes from); while this is not always true, the technique is still surprisingly effective.

5. Usually, the “non-events” – that is, the non-appearances of a feature in a text – are also considered, which leads to a straightforward extension of Eq. 2.

## 4 Applying the Naive Bayes Method to Image Classification

The first problem hindering the application of naive Bayes to image classification is: what are the analogues of “text” and “feature” in images? For the task

of detecting images which contain some pre-defined structures, one may define a feature as a certain sub-image. For example, for detecting images with human faces, a useful feature would be the presence of an eye in the image. In [20], such “informative features” were recovered, and various algorithms used to classify images based on the features’ presence. Certain textures can also be recognized by the presence of templates, perhaps up to rotation or scale, etc. Such features, however, are unsuitable for the problem of style detection as presented here (unless we identify a painting by the artist’s signature...). In general, one cannot hope to base “style classification” on the presence of a few features.

Instead, we offer to classify every image block, and then classify the entire image by a majority vote. The information extracted from this process contains more than the classification of the entire image; it maps the image to different regions, each dominated by a certain style. As will be demonstrated in Section 7, this often yields results which agree with human intuition. The local analysis of the image contains more information than that present in histogram-based approaches, which classify the entire image based on similarity between cumulative distributions of gradients, or wavelet coefficients etc.

As opposed to the text categorization applications of naive Bayes, and also to [20], this paper suggests to use *features which have the same size as texts*. We treat each and every image block (the size in the experiments was  $9 \times 9$ ) as a text, and the features are the  $9 \times 9$  DCT basis functions. We say that a certain such feature (coefficient) appears in a block if its absolute value in the block’s expansion is larger than a certain threshold; in Section 5 we explain how this threshold is determined (see also Fig. 2).

## 5 Implementation

The suggested implementation of the classifier to the problem of “style” detection proceeds as follows (the explanation is presented to classification of paintings, but

the algorithm is general):

1. Build an image database. Here, we have tested five artists – Rembrandt, Van-Gogh, Picasso, Magritte, and Dali. Ten paintings by each artist consisted the training set, and the test set consisted of twenty to thirty paintings for each artist. The training set was randomly chosen.
2. For each DCT basis element ( $9 \times 9$  in size),  $b_{ij}$ , and for every artist, the absolute values of the DCT coefficient corresponding to  $b_{ij}$  are computed for every  $9 \times 9$  block in all the artist’s paintings in the training set. These values are then binned into 100 discrete values. The blocks are first normalized to zero mean and unit variance, hence the absolute values of the coefficients are between 0 and 1. All these operations can be implemented using convolutions, hence can be done rather quickly. Then, it is straightforward to construct a table  $T(p, i, j, a)$ , which stores the probability that, for the artist  $p$ , the absolute value of the  $(i, j)$  DCT coefficient is greater or equal than  $a$ . Here  $p$  ranges over all artists,  $i$  and  $j$  between 1 and 9, and  $a$  ranges over  $\{0, 0.01, 0.02, \dots, 0.99, 1\}$ .
3. Naive Bayes requires binary features, so we have to convert the continuous presence of a basis element in a block (that is, its coefficient in the block’s expansion), to a binary one. This is done by thresholding the coefficient’s absolute value. For every pair of artists and every coefficient, the threshold is chosen so as to maximize the mutual information (Eq. 1). Note that this is a very fast process, once the probability table of stage 2 was built. The maximization is performed over each binned value  $\{0, 0.01, 0.02, \dots, 0.99, 1\}$ , and over both artists.
4. For each artist in each pair, the ten features with the highest mutual information are chosen. Note that each feature consists of a basis element and a threshold for its coefficient in a block’s expansion.

5. Given a new image and a pair of artists, the probability of each image block with respect to each artist is computed from Eq. 2. Better results were obtained by considering only blocks with a (pre-normalization) variance higher than a certain threshold – 20 was a good value, but the results don’t change much if 10 or 30 is used. Another heuristic which yielded better results was to classify only blocks for which the winning artist’s probability was at least twice the other artist’s probability.
6. Every pixel in the test image is assigned a label, according to the classification of the  $9 \times 9$  block surrounding it. Pixels whose corresponding window’s variance is too small, or for which the ratio between the large and small probabilities does not exceed 2, are labeled as unclassified.
7. The overall classification is determined by a majority vote. However, as noted before, the mapping of individual pixels to different artists contains more information than the overall classification.

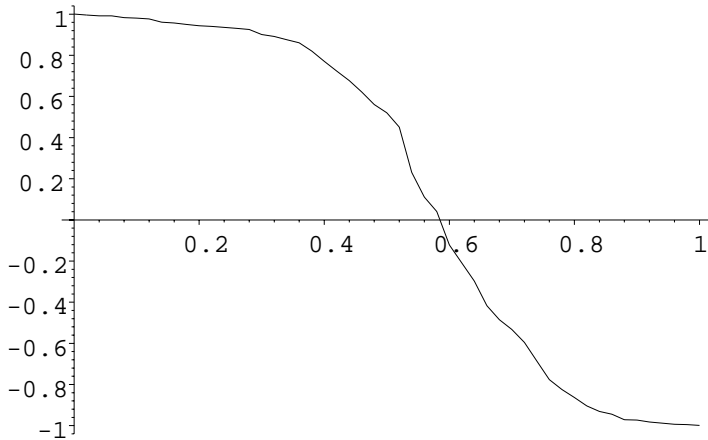
## 5.1 Why DCT?

Recall that naive Bayes assumes that the features are independent. Since the DCT basis elements are orthogonal, if  $(i_1, j_1) \neq (i_2, j_2)$ , and  $b_{i_1, j_1}, b_{i_2, j_2}$  are the respective basis elements, then under the “natural” probability distribution over images [11], the random variables  $I \rightarrow (I, b_{i_1, j_1})$  and  $I \rightarrow (I, b_{i_2, j_2})$  (where  $I$  varies over images), are independent random variables over the space of *all* images.

One may rightly question the validity of the independence of the DCT coefficients, which may be false in a “small world”, corresponding to a single artist for example. Therefore, a decorrelated basis was created by diagonalizing the covariance matrix for each artist, and classification according to this basis was attempted. Somewhat to the author’s surprise, this didn’t result in any substantial improvement. A possible explanation as to why naive Bayes can overcome feature dependencies is offered in [12].

## 6 Classifying a Mixture of Styles

Suppose we are given a mixture (or average) of images of different styles. This can take place, for example, when transparencies exist in the scene (e.g. looking into a room via a window in which greenery is reflected). It may be expected that, at least as far as human intuition is concerned, the classification of the mixture will not be linear; that is, the classification of a weighted average of styles will be biased towards the dominant one. To test the behavior of the classifier on a mixture of styles, weighted averages of two paintings – a self-portrait by Van-Gogh ( $P_1$ ) and “Metamorphosis of Narcissus” by Dali ( $P_2$ ) – were created (after the images were normalized to the same size). The mixtures varied over convex combinations,  $\lambda P_1 + (1 - \lambda)P_2$ , for  $\lambda = \frac{k}{100}$ ,  $0 \leq k \leq 100$ . For each  $\lambda$ , the mixture image was classified. A numerical measure assigned to each  $\lambda$  was defined as the difference between the number of “Dali blocks” and “Van-Gogh blocks”. The two “pure” images ( $\lambda = 0, 1$ ) were normalized to 1 and  $-1$  respectively. Fig. 1 depicts the relation between  $\lambda$  and this measure:



**Please note: this figure is in color.**

Figure 1: Left: Strength of classification as a measure of  $\lambda$  (which is represented by the horizontal axis), of the mixture  $\lambda P_1 + (1 - \lambda)P_2$ , where  $P_1$  stands for Van-Gogh and  $P_2$  for Dali. Classification strength is determined by the ratio of  $P_1$  vs.  $P_2$  blocks. The strength is linearly normalized such that the cases  $\lambda = 0$  (“pure Dali”) and  $\lambda = 1$  (“pure Van-Gogh”) are assigned strengths of 1,  $-1$  respectively. Results are consistent with human intuition which is biased towards the dominant style. This non-linear behavior, which results in the classification being robust under “contamination” by another style, is a result of the non-linear thresholding of the DCT coefficients. Right: one of the images used in the mixture classification. “Dali pixels” are reddish, Van-Gogh bluish. Here and elsewhere, unclassified pixels are in gray level.

## 7 Results

For the five artists tested, a “tournament scheme” classifier was implemented [16].

The rate of success was 86%. Some examples are presented below.

**Please note: Figs. 3-7 are in color.**

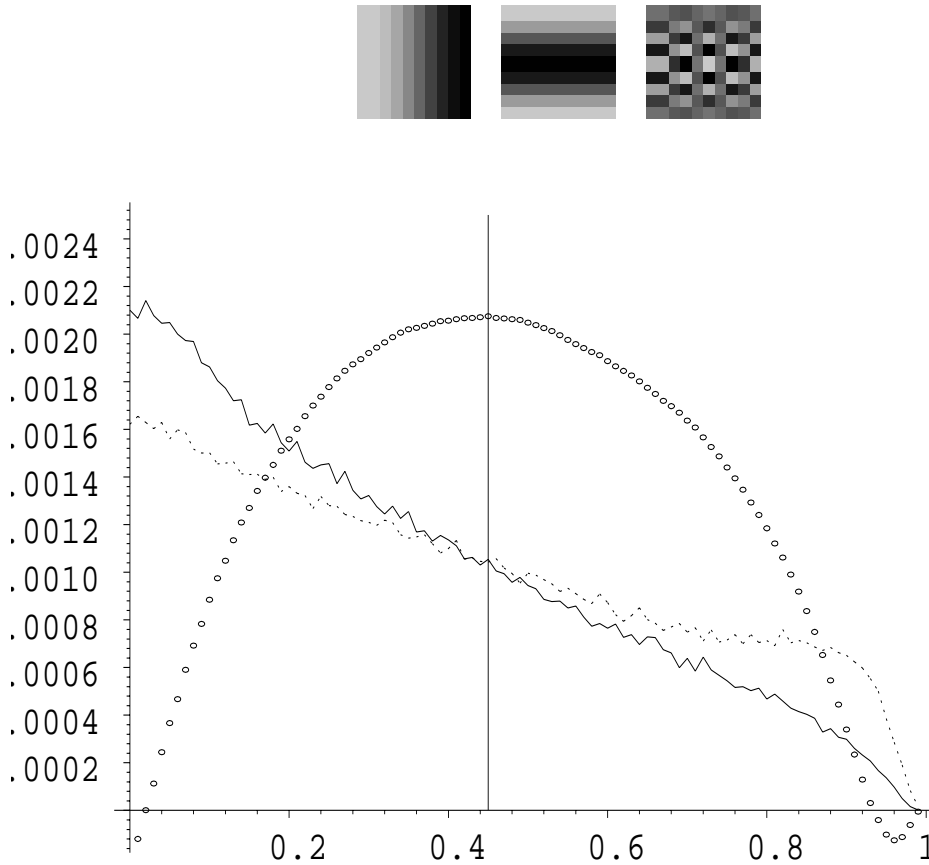


Figure 2: Top: three DCT basis elements with highest mutual information for discriminating Dali from Van-Gogh. The mutual informations (left to right) were 0.042, 0.037, 0.036. Bottom: the distribution and mutual information for various thresholds of the DCT coefficient corresponding to the most discriminating feature (top left). The mutual information for thresholds between 0 and 1 is the parabolic-like curve (empty circles); it is scaled by 0.05 for visualization purposes. The percentage of Dali blocks with the corresponding DCT coefficient (after binning) is the dotted line, and the solid line depicts the same for Van-Gogh blocks. The optimal threshold (for which the highest mutual information is obtained) is 0.45 (solid vertical line). Note that it is achieved roughly at the point in which the Dali and Van-Gogh curves intersect.

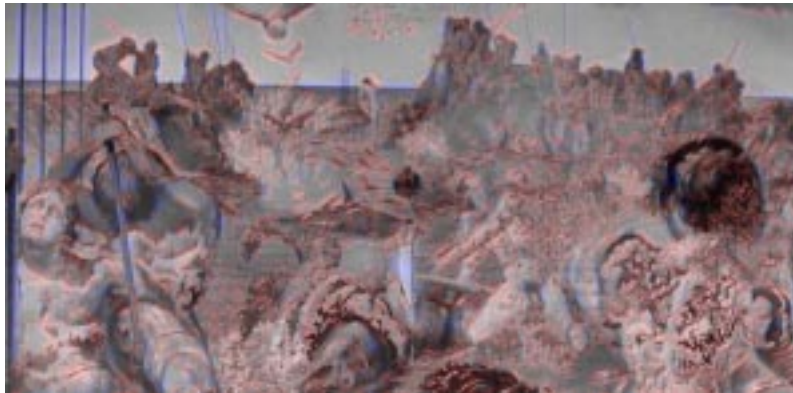
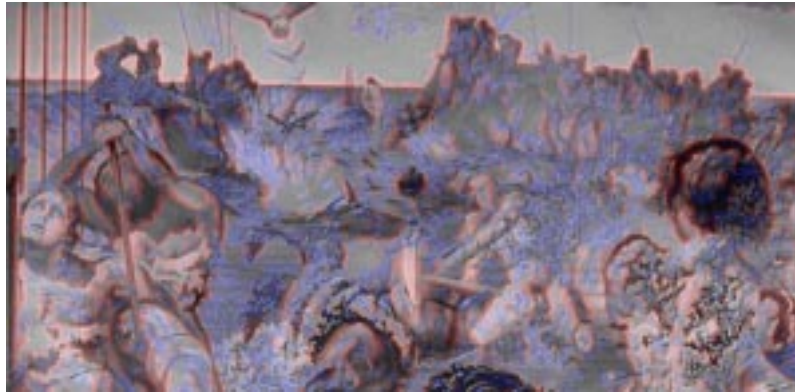


Figure 3: Top: excerpt from Dalí's "Tuna Fishing". Here and elsewhere, the classification was performed after the painting was transformed to gray levels. Middle: same excerpt after Dalí/Van-Gogh classification; "Dalí pixels" are reddish, "Van-Gogh pixels" bluish. In accordance with human intuition, "wavy" areas are dominantly Van-Gogh (for example, the part of the arm at the bottom right corner). Bottom: same excerpt after Dalí/Magritte classification; "Dalí pixels" are reddish, "Magritte pixels" bluish. Note, for example, that the vertical sharp structures in the top left are classified as Dalí when compared against Van-Gogh, but as Magritte when compared against Magritte. This can be intuitively explained by the observation that "Dalí paints with more straight lines than Van-Gogh but with less straight lines than Magritte".



Figure 4: Top: The painting “Wheat Field with Cypress” by Van-Gogh. Bottom: after classification vs. Dali. Van-Gogh pixels are reddish, Dali pixels bluish. The typical Van-Gogh wavy structures are correctly classified in general.

## 7.1 Classifying Images as “Old” or “New”

Some experiments were made in discriminating old photographs (19th century) and photographs captured by a digital camera; the results are displayed in Fig. 5.



Figure 5: Top: Four of the images used as a training set for “old pictures”. Bottom: Result of classification for a photograph of Robert E. Lee. Pixels classified as “old photograph” pixels are reddish, “digital camera” are bluish.

## 8 Activity Detection in Video

The “style classification” method was extended to detect activities in video sequences. This topic is drawing a lot of interest in recent years [4, 8, 5]. The literature is too numerous to cover in this short paper; for a good survey, one may consult [13]. Temporal texture and gradient distribution approaches, which can be viewed as an extension of texture classification in the spatial domain, are discussed in [14, 15, 23].

In order to recover features which characterize a certain type of activity, the algorithm in Section 5 was extended to the spatio-temporal domain. First, three-dimensional “image stacks” are built from the movie segments in the training set. This is accomplished by constructing a three-dimensional array  $A$ , with  $A[i, j, k]$  = the  $(i, j)$  pixel in the  $k$ -th frame. In order to save memory, this can be done sequentially, with the probability table (stage 2 in Section 5) built by adding the data for sub-sequences. Instead of two-dimensional blocks, three-dimensional blocks are used, with the 3D DCT transform coefficients as features. A block  $B$  corresponding to the pixel with spatial coordinates  $(x, y)$  at time  $t$  consists of the pixels  $B[i, j, k] = A[x + i, y + j, t + k]$ , where  $i, j, k$  range in some small intervals centered at 0 (we used  $5 \times 5 \times 5$  blocks). Blocks with a small time derivative (i.e. in which not much activity occurs) are not considered. This limits the algorithm to a stationary camera. If the camera is moving, stabilization and motion compensation can be used. The experiments were so far limited to distinguishing between two types of activity: hand waving and walking. Four different individuals were filmed while performing these activities, in three different locations: opposite a white wall, in an office, and in a corridor. Part of the captured video sequences were used for training, and then the classification algorithm was applied to the remaining sequences. A resolution of  $64 \times 64$  was used, which is low enough to allow real-time classification.



Figure 6: Eight frames from a low-resolution video sequence depicting a person walking across a corridor. If the spatio-temporal  $5 \times 5 \times 5$  neighborhood of a pixel was classified as “walking” it was colored purple, and if it was classified as “hand waving” it was colored yellow. Most of the misclassification occurs in areas in which the diagonal motion of the legs resembles the upwards or downwards motion of the hands in the “hand waving” sequence (see Fig. 7). Some of the person’s reflection is also classified as ”walking”. Altogether, 83% of the classified pixels were labeled as “walking”.

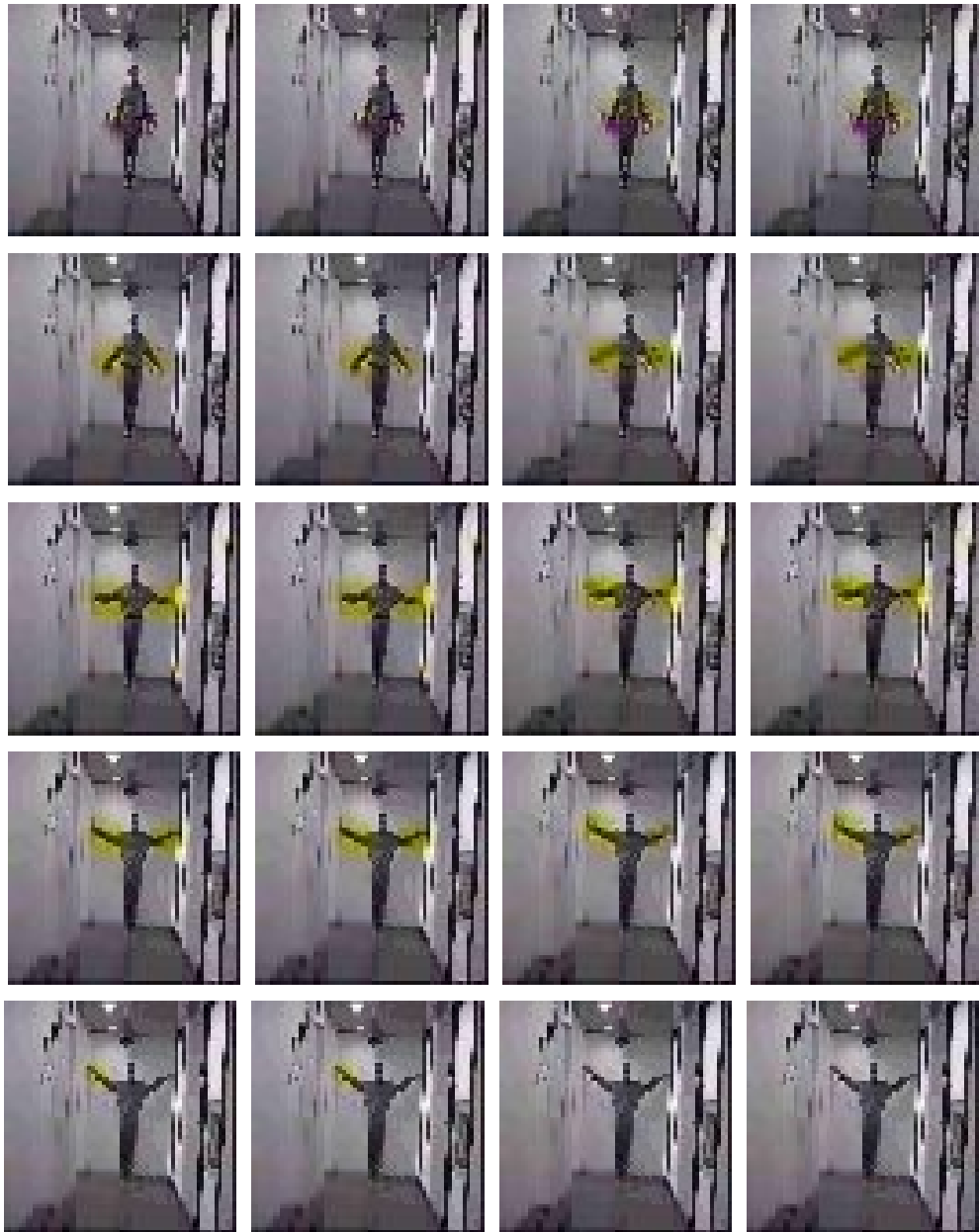


Figure 7: Twenty frames from a low-resolution video sequence depicting a person waving his hands. If the spatio-temporal  $5 \times 5 \times 5$  neighborhood of a pixel was classified as “walking” it was colored purple, and if it was classified as “hand waving” it was colored yellow. Most of the misclassification occurred within the first four frames, in which the diagonal motion of the hands at a low inclination locally resembles the leg motion of a walking person. Barely any motion was detected in the last four frames, because they depict the stage in which the upwards motion of the hands ends and there’s a slight pause before the downwards motion commences. Altogether, 98% of the classified pixels were labeled as “hand waving”.

## 9 Conclusion and Future Research

A simple and very fast algorithm for image and activity “style” classification using the naive Bayes classifier was presented, and applied to the problems of artist identification, classifying photographs as digital or old, and activity detection in video (walking vs. hand waving). Further research will consist of incorporating a multi-level scheme and developing methods to determine the correct block size(s), as well as testing other representations than the DCT, such as wavelets and overcomplete bases. A Markov random field paradigm may be applied in order to create a more consistent segmentation of the image (i.e. not to allow an isolated “Dali pixel” in a “Van-Gogh area” of the image, as well as to preserve continuity of action over time), as well as additional high-level processing and grouping.

The advantages of the presented algorithm are in its simplicity and speed, and its ability to handle a large number of features. The results are reasonable and consistent with human intuition. Reasonable results were also obtained for video sequences although the resolution was rather low.

The suggested method is very local in nature and thus can handle a few different styles or activities which co-exist in an image or a video sequence. It can also overcome considerable “contamination” of one style by another, as demonstrated in Section 6.

## References

- [1] K. Barnard and D.A. Forsyth. Learning the semantics of words and pictures. In *International Conference on Computer Vision*, pages II: 408–415, 2001.
- [2] M. Bell and W.T. Freeman. Learning local evidence for shading and reflectance. In *International Conference on Computer Vision*, pages I: 670–677, 2001.

- [3] S. Belongie, C. Carson, H. Greenspan, and J. Malik. Color- and texture-based image segmentation using the expectation-maximization algorithm and its application to content-based image retrieval. In *International Conference on Computer Vision*, pages 675–682, 1998.
- [4] M.J. Black, Y. Yacoob, A.D. Jepson, and D.J. Fleet. Learning parameterized models of image motion. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 561–567, 1997.
- [5] A.F. Bobick and J.W. Davis. The recognition of human movement using temporal templates. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(3):257–267, March 2001.
- [6] J.S. de Bonet and P.A. Viola. Texture recognition using a non-parametric multi-scale statistical model. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 641–647, 1998.
- [7] S. T. Dumais, J. Platt, D. Hecherman, and M. Sahami. Inductive learning algorithms and representations for text categorization. In *Proc. 7th International Conference on Information and Knowledge Management*, pages 148–155, 1998.
- [8] D.J. Fleet, M.J. Black, Y. Yacoob, and A.D. Jepson. Design and use of linear models for image motion analysis. *International Journal of Computer Vision*, 36(3):169–191, February 2000.
- [9] M.D. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker. Query by image and video content: The qbic system. *Computer*, 28(9):23–32, September 1995.

- [10] H. Greenspan, R. Goodman, R. Chellappa, and C.H. Anderson. Learning texture-discrimination rules in a multiresolution system. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 16(9):894–901, September 1994.
- [11] D. Keren and M. Werman. Probabilistic analysis of regularization. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 15:982–995, October 1993.
- [12] D. Lewis. Naive bayes at forty: The independence assumption in information retrieval. In *Proc. 10th European Conference on Machine Learning*, pages 4–15, 1998.
- [13] T.B. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 81(3):231–268, March 2001.
- [14] R. Polana and R.C. Nelson. Detecting activities. In *International Conference on Pattern Recognition*, pages A:815–818, 1994.
- [15] R. Polana and R.C. Nelson. Temporal texture and activity recognition. In *Motion-Based Recognition, Chapter 5*, 1997.
- [16] M. Pontil and A. Verri. Support vector machines for 3D object recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(6):637–646, 1998.
- [17] E.P. Simoncelli and E.H. Adelson. Noise removal via bayesian wavelet coring. In *IEEE International Conference on Image Processing*,, pages I: 379–382, 1996.
- [18] M.G.A. Thomson and D.H. Foster. Role of second-order and third-order statistics in the discriminability of natural images. *Journal of the Optical Society of America, A*, 14(9):2081–2090, September 1997.

- [19] K. Tieu and P. Viola. Boosting image retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages I:228–235, 2000.
- [20] S. Ullman, E. Sali, and M. Vidal-Naquet. Fragment-based approach to object representation and classification. In *Visual Form 2001, 4th International Workshop on Visual Form, IWVF-4, Capri, May 2001*, volume 2059 of *Lecture Notes in Computer Science*, pages 85–102. Springer, 2001.
- [21] R.F. Voss. Local connected fractal dimension analysis of early chinese landscape paintings and x-ray mammograms. In *Fractal Image Encoding and Analysis: A NATO ASI Series Book*, pages 279–297, 1996.
- [22] J.Z. Wang, J. Li, and G. Wiederhold. Simplicity: Semantics-sensitive integrated matching for picture libraries. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(9):947–963, September 2001.
- [23] L. Zelnik-Manor and M. Irani. Event-based analysis of video. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages II:123–130, 2001.