14. International Conference on Pattern Recognition, August 17-20, 1998, Brisbane, Australia

Sample-guided Progressive Image Coding

Udo Mahlmeister, Michael Sandgaard, and Gerald Sommer
Christian–Albrechts–Universität zu Kiel
Preußerstraße 1–9, D–24105 Kiel, Germany
email: uhm@informatik.uni–kiel.de

Abstract

We introduce a low-level image representation using local orientation on a steerable pyramid. Its performance is demonstrated for a purposively progressive image coder which transmits and reconstructs parts of an image guided by their similarity to a given sample image. Like human visual attention, the image is scanned in a most-important-first order to gain a maximum of visual information with minimum code.

1. Introduction

Browsing in a remote image database is often hampered by the considerable delay of transmission. As long as this delay is noticed by the user, he would like to abort the transmission of the current image and skip to the next when his task (e.g. recognizing a face) is just accomplished or – perhaps the more interesting case – is not expected to be accomplished with additional information. If the relevant part of visual information is limited to a small part of the image (e.g. the eyes of a face) and located at the beginning of the code this method could substantially save time and transmission costs. This idea is illustrated by the reconstruction sequences in fig.1 and 8 which are observed at the receiver. In a strategy called Purposively Progressive Image Coding (PPIC), we propose to re-arrange image parts according to their relevance for the user's task. By sending most relevant parts before others, transmission could be aborted as early as possible whereas delivering a maximum of relevant information. PPIC is strongly inspired by visual attention which enables humans to purposefully scan a scene by means of proper eye movements.

A successful technical realization of PPIC depends on four factors: 1) a precise task specification by the user, 2) an efficient and robust localization of the visual information relevant to this task, 3) a partial but perceptually complete reconstruction, and 4) a sparse code thereof. Whereas there are some results for 3) and 4), there are barely solutions to 1) and 2), because they are considered unrelated to 3) and 4). This has also been realized by some authors[5, 9], who sketch Visual Information Management Systems (VIMS) for browsing image databases. These systems are intended to purposively stratify the images in a data base for a limitation of the users search space. Analogously we propose a PPIC to purposively *stratify the parts of an image* to minimize the number of transmission steps for a given task. Using an image representation common to the four points is not merely to minimize time consuming low level operations[9]. As we will show in the rest of this paper, it is the key issue of PPIC.

Figure 1. Reconstruction sequence. Left to right: a) 4th ~ 2.5%, b) 12th ~ 7.5%, c) last step from 158.

In the next section we introduce the local orientation pyramid (LOP) as this common representation. Section 3 discusses the localization and coding aspects of LOP. In section 4 we provide results from experiments with a prototype PPIC system applied to face images.

2. Local Orientation Pyramid

To provide the user with an expressive language for a convenient description of his task (point 1) a representation should support an effortless translation to and from real world concepts, e.g. orientation, contrast, color. Though this point is important for VIMS we do not address it here. Instead, we confine our discussion on *progression by sample* which requires the user to provide a small sample image. Then the system partitions and transmits image parts in a best-match-first order. For this purpose we need to define a distance between an image part waiting for transmis-

sion and the sample image. This distance has to be robust to variations of viewing angle, illumination intensity, and color to guarantee a reliable localization of relevant image parts. Robustness to these image formation parameters is gained by two unrelated mechanisms. First, local orientation is extracted at multiple scales. It provides robustness to photometric distortions since the component of illumination in the signal is suppressed by bandpass filtering[7]. Features with further, especially geometrical, invariants have attracted major interest in pattern recognition theory. However, they are not desirable here since their complexity of computation makes human interaction difficult. Furthermore, they are too lossy for a stable discrimination of similar objects. In order to achieve robustness to additional parameters, we propose local feature histograms as the second mechanism[13]. Similarity between histograms of simple features have proved to be computationally very efficient and robust to geometrical distortions.

For an efficient coding architecture we would like to use the same representations for localization and reconstruction. However, there are contradicting requirements arising from the two problems. Localization relies on useful redundancy in the signal in order to gain robustness to distortions – the basic idea of error correcting codes. As opposed to this, in image compression the aim is to remove redundancy by concentrating most of the information on few active elements. This is Barlow's principle of sparse coding. By using the steerable pyramid we meet both requirements since we may control how much redundancy the representation contains.

The *Local Orientation Pyramid (LOP)* as introduced in the sequel is based on the steerable pyramid[11]. It provides joint steerability in orientation and position. Steering position serves to interpolate missing samples in the pyramid, that is to convert the pyramid into a heap. Thus, inter-scale vectors may be constructed at the lowest pyramid level, providing highest resolution for localization. Steerability in orientation, on the other hand, is required for calculating local amplitude/orientation as shown below.

The block diagram of the steerable pyramid architecture is depicted in fig.2. By replacing the filled dot with the dashed box a new level of the pyramid is constructed. Four pyramid levels are used with K orientation selective filters $\mathcal{B}_i$, $i = 0 \ldots K-1$ each. The corresponding kernels $b_i(x,y)$ are designed as rotated copies at orientations $\Theta_i = i \cdot 180°/K$. Thus a filter $b^{(\Theta)}(x,y)$ at arbitrary orientation Θ may be synthesized by the linear combination[2]

$$b^{(\Theta)}(x,y) = \sum_{i=0}^{K-1} k_i(\Theta)\, b_i(x,y) \qquad (1)$$

Using $K = 4$ filters at orientations $\Theta_i = i \cdot 45°$ we obtain interpolation functions[2]

$$k_i(\Theta) = \frac{1}{2}\left(\cos(\Theta - \Theta_i) + 2\cos 3(\Theta - \Theta_i)\right) \qquad (2)$$

Because the limited spectral support of the filters is not exploited for further sub–sampling, the representation is overcomplete[12, 11] by a factor $16/3$.

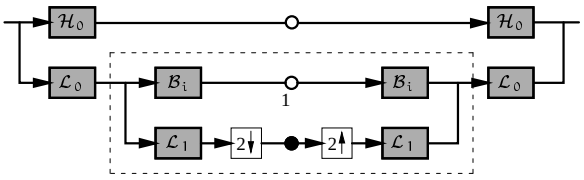

Figure 2. Steerable pyramid[11]: initial high-pass $\mathcal{H}_0$, low-pass $\mathcal{L}_0$, recursive subsystem (dashed) with low-pass $\mathcal{L}_1$, sub-sampling, and orientation selective band-passes $\mathcal{B}_i$.

In order to convert the pyramid into a heap the outputs of the $\mathcal{B}_i$ at each level have to be up-sampled by a factor 2. This is achieved by inserting an up-sampling operator and a low pass filter $\mathcal{L}_1$ at position 1 of fig.2. The $\mathcal{L}_1$ serves to interpolate the missing samples, that is to steer position.

Besides continuous position and orientation a continuous scale space could be constructed by steering scale. Actually, this approach seems theoretically quite appealing since objects at arbitrary viewing distances appear at continuous scale. However, there are two arguments against steerable scale in PPIC. First, it is numerically considerably more complex than steering position or orientation due to the singularity at scale infinity[8]. Second, the energy of most image patterns is distributed over multiple scales. This requires inter-scale feature vectors for representation which in turn require a discrete scale. Probably these constraints are also valid for mammalian visual systems which do not steer scale either[1]. Orientation behaves completely different. As indicated in fig.3, at most one orientation is sufficient for the description of most frequent patterns (plan and linear) in natural images. Inter-orientation vectors which are used in sub-band coders are therefore redundant. We prefer the *single orientation model* which provides a more concise representation for classification and reconstruction.

Figure 3. Plan(l), linear(m), complex(r) shaped points.

The single orientation model assigns to each point of the image a local amplitude a and a local orientation ϑ which are related to the contrast and the angle of predominant orientation of the pattern, respectively[4]. Given the K outputs h_i of LOP filters $\mathcal{B}_i$ we define:

$$a := \left| \sum_{i=0}^{K} |h_i| \, e^{-j2\Theta_i} \right| \qquad \vartheta := \arg \sum_{k=0}^{K} h_i \, e^{-j\Theta_i} \qquad (3)$$

Closely related with local amplitude/orientation is the concept of local amplitude/phase A and φ. Its motivation is to decouple the detection and classification of oriented patterns by segregating contrast and shape[4]. As explained in the following, this strategy is controversial if the statistics of natural images is considered. Disregarding A the marginal distribution of φ is flat (due to its linearity). Thus, the role of φ as an own feature dimension is justified. However, the correlation between φ and the symmetry of a pattern is valid just at the center of symmetry which coincides with the maximum of A. If φ is considered exclusively there its distribution shows peaks at $\pi/2$ and $3\pi/2$ – the phase values of positive and negative odd symmetric edges, respectively. Other values of φ are mostly faked due to a dislocated maximum of A. To suppress this irrelevant information, we exclusively use odd symmetric filters in (3). The corresponding phase values $\pi/2$ and $3\pi/2$ are represented by local orientation intervals $[0, 180°)$ and $[180°, 360°)$. Thus, our local orientation as defined in (3) spans the full circle $[0, 360°)$ as opposed to Granlunds[4] which spans the half circle, only. The implications of this definition for the local orientations of a square are depicted in fig.4.

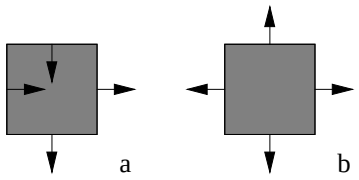

Figure 4. Local orientation defined for a range of value $[0, 180°)$ in (a) and $[0, 360°)$ in (b).

3. Localization and Coding with LOP

Progression by sample requires an image to be partitioned into spatially independent components which are sorted in decreasing similarity to the sample image. A robust similarity metric as claimed in section 2 is realized by two operators with different region of support. Illumination changes are suppressed by the lateral inhibition of the LOP filters. They realize a local Retinex–like[6, 3] mechanism for an approximate color constancy. Additional insensitivity to illumination is achieved by a coarse quantization of local amplitude which carries the contrast information. Thus, the color constancy of our method is superior to those using filter responses directly[10, 3]. Robustness to geometrical distortions is achieved by the larger local context of the sliding histogram window whose size is adapted to the size of the sample image. Though rather sensitive to changes of illumination, histogram similarity of RGB-vectors has been proved to be robust to geometrical distortions i.e. changes of scale, rotation, viewing angle and even partial occlusion[13, 3, 10]. This property is successfully combined with color constancy if histograms are calculated from inter-scale local orientation vectors. These vectors are constructed from three quantized pairs (a, ϑ) of the first three levels of LOP. Quantizers contain five code vectors each: a "null orientation" which is assigned to patterns with contrast below a certain perceptual threshold a_{thr}, and another four code vectors corresponding to patterns with significant contrast at orientations $\vartheta = 0°, 90°, 180°, 270°$. In fig.5a the quantization for the complex plane of $z = a(\cos\vartheta + j\sin\vartheta)$ is illustrated. Code vectors are represented by grey values.

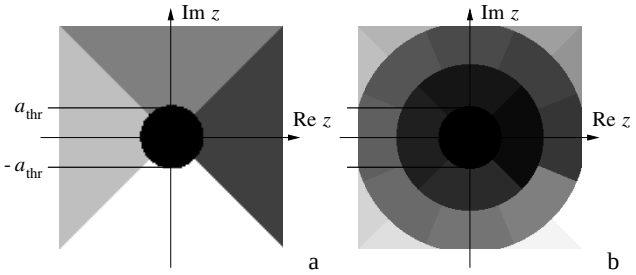

Figure 5. Quantization of local orientation $z = a(\cos\vartheta + j\sin\vartheta)$ using L code vectors (represented by grey values), for (a) localization $L = 5$, (b) coding $L = 14$.

In the following we describe four distance metrics (`INT`), (`CHI`), (`MAP`), and (`BPR`) between the sample image and each local neighborhood of the input image by the using quantized orientations described above. The former two methods involve the calculation of histograms M_ℓ and I_ℓ, $\ell = 1 \ldots L$ of the sample image and a window scanning the input image. Using three LOP-levels with five code vectors each the number of histogram bins amounts to $L = 5^3 = 125$. The distance between M_ℓ and I_ℓ calculated by (`INT`) or (`CHI`) yields an estimate for the similarity between the corresponding images. Histogram intersection (`INT`) as defined in [13] is said to need sparse histograms for sufficient discrimination performance[13, 10].

$$\eta(I, M) = \sum_{\ell=1}^{L} \min(I_\ell, M_\ell) \qquad (\texttt{INT})$$

Since histograms of multi–scale orientation vectors are hardly sparse, we apply the χ^2–test (`CHI`) as an alternative metric. The value of (`CHI`) indicates how different two histograms I and M (supposed to be distributed normally) are:

$$\chi^2(I, M) = \sum_{\ell=1}^{L} \frac{(I_\ell - M_\ell)^2}{I_\ell + M_\ell} \qquad (\texttt{CHI})$$

Sliding histogram techniques are rather complex if they are performed in parallel. On a sequential computer, however,

a meandering sampling leads to a very efficient updating strategy. The following two methods completely avoid to calculate the histogram of a sliding window. Based on the global histograms of image and sample they require the application of a point operator, followed by a low pass filter whose support is determined by the size of the histogram window. The first of these methods applies Bayes' formula(MAP) to yield the maximum a–posteriori probability that a specific position belongs to an object O if a feature F_ℓ is present there.

$$p(O|F_\ell) = \frac{p(F_\ell|O)p(O)}{p(F_\ell|O)p(O) + p(F_\ell|B)p(B)} \qquad \text{(MAP)}$$

The probabilities $p(F_\ell|O)$ and $p(F_\ell|B)$ are to the normalized feature histograms of the sample object O and the background B. The a–priori probabilities $p(O)$ and $p(B)$ correspond to the relative area occupied by O and B in the image. For a single object $p(B)$ equals $1 - p(O)$.

The other method, histogram back-projection as defined in [13] is a straight simplification of (INT).

$$R_\ell = \min\left(1, \frac{M_\ell}{I_\ell}\right) \qquad \ell = 1 \ldots L \qquad \text{(BPR)}$$

It is also a special case of (MAP), since M_ℓ and I_ℓ are estimates of $p(F_\ell|O)p(O) = p(F_\ell \cap O)$ and $p(F_\ell|O)p(O) + p(F_\ell|B)p(B) = P(F_\ell)$ for the current image.

In contrast to the rest the Bayes formula (MAP) includes the additional quantities $p(O)$, $p(B)$, and $p(F_\ell|B)$. Whereas there are no problems in estimating the former two, errors in the estimation of the background histogram $p(F_\ell|B)$ induce misleading biases for localization. In addition, these quantities have to be determined at query time.

Whereas the pairs (a, ϑ) are quantized quite coarsely for localization, the reconstruction from sub-LOPs requires a finer quantization which is specificly tuned to perceptual resolution of each pyramid level. As in the former case, a dead zone $a < a_{thr}$ representing the null orientation accounts for contrast threshold observed in the human visual system. Its size is set to a small value for the highest pyramid level and to a large value for the lowest level. As illustrated in fig.5b, the number of orientation bins is coupled to the number of amplitude levels and the actual amplitude level. This procedure accounts for the phenomenon that the orientation of high contrast patterns is resolved more accurately by humans than those of low contrast patterns. Note that unlike quantizing coefficients of orthogonal wavelet transforms our quantization scheme doesn't induce any directional preferences thus enabling real 2D-quantization. The low–pass $\mathcal{L}_1$ of the last LOP-level (see fig.2) is quantized separately by a scalar quantizer. It is sent first to give the user a coarse impression of the image. The remaining high–pass $\mathcal{H}_0$ is discarded.

Besides the quantized bands the positions of sub-images have to be transmitted. This side information could amount a substantial part of the transmission costs if progression has to be available until the final reconstruction is reached. But it is kept quite low ($\approx$ 5bit per progression step) if progression is restricted to e.g. a quarter of the image.

4. Experiments

The performance of progression by example depends on the graceful degradation of histogram similarity for distortions of the image with respect to the model. Ideally, the sequence of reconstruction should not be corrupted by such distortions. Each of the four histogram metrics requires particular constraints on the histogram (e.g. sparse, normal[13, 10]) in order to work robust. These constraints can be guaranteed neither off-line, lacking a complete statistics of the quantized coefficients for natural images, nor online for reasons of coder efficiency. Therefore, we evaluate the discrimination performance of each method for the test image in fig.1c and its distorted versions fig.6`pan`–`texture`.

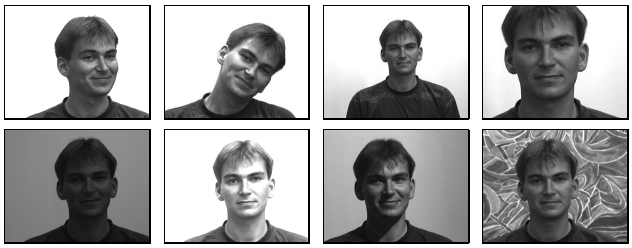

Figure 6. Test images. From left to right: `pan`, `roll`, `small`, `large`, `dark`, `bright`, `shady`, `texture`

As is well known, the aim of image coding is to reduce the data rate while maximizing the image fidelity with respect to the original. Fidelity is usually measured in terms of signal-to-noise ratio (SNR) which is obviously inappropriate for PPIC, since it ignores the unequal relevance of image components for a specific task (e.g. the eyes in a face for face recognition). Fortunately, progressive coding itself provides an interactive method to evaluate PPIC sequences. Presented with an equally partitioned sequence, a user is asked to stop the process of reconstruction when his task is accomplished. The number of image components reconstructed so far is called *number of steps to succeed* (STS). Given the set of all possible sequences obtained by a permutation of components, the sequence with the minimal STS is optimal since contains the most relevant image components within the first STS steps. Nevertheless, the STSth reconstruction step need not to be close to the original in terms of SNR but it provides the highest compression ratio with respect to the users task. For the particular task of face recognition with LOP in particular, we define STS as the number of sub-pyramids required for recognizing the eyes as in fig.1b.

In our experiments we examined to what extent the STS

match	pan	roll	small	large	dark	bright	shady	texture	norm	monkey	black boy	Asian girl	glasses	up-sample
CHI	26	26	26	27	28	18	18	31	17	20	38	15	19	23
INT	15	15	18	17	13	9	7	15	8	9	27	6	6	11
BPR	5	21	22	27	36	7	9	30	10	23	30	14	15	20
MAP	20	47	46	37	37	4	16	78	12	77	31	39	28	27
	a								b	c				d

Table 1. Number of steps to succeed (STS)

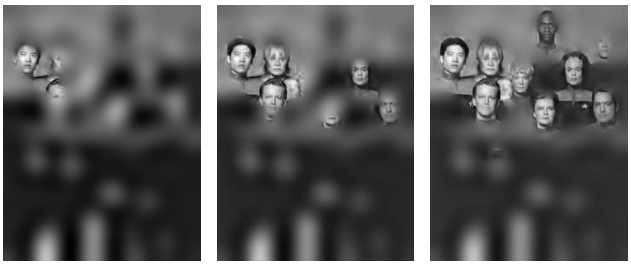

Figure 8. Reconstruction sequence. Left to right: a) 5th $\sim$ 2.5%, b) 20th $\sim$ 10%, c) 45th $\sim$ 23% step from 196 using a face as sample image.

is affected if test or model image deviate from the original. Two cases are of practical importance. First, the input image differs from the model in perspective transformation or illumination. For this test the eight distorted test images of fig.6 were reconstructed using the original eye in fig.7`original`. In tab.1a the corresponding STS's are listed in comparison to the STS for a reconstruction of the original portrait fig.1c. To sum up it can be said that `INT` outperforms its competitors. Though superior for `pan` and `bright` the local methods `BPR` and `MAP` have to be regarded as unreliable, due to their unsteady performance.

In the second practically important case the model image is taken from an arbitrary sample of the object class describing the users task. In particular, we consider the reconstruction sequences of the original face image (fig.1c) with other peoples eyes (fig.7`monkey–glasses`) as models. The resulting STS's in tab.1c confirm that `INT` is clearly superior. In the last experiment we show that steering position improves progression. The STSs decrease by 26–55% if steering position (tab.1b) is substituted by a simple up–sampling (tab.1d). Fig.8 indicates to the application of PPC on more complex models like faces.

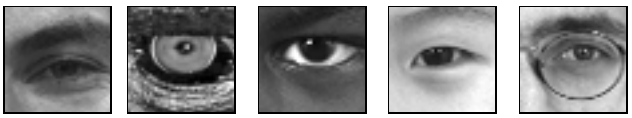

Figure 7. Model eyes `original` (from fig.1c), `monkey`, `black boy`, `Asian girl`, `glasses`.

5. Conclusions

We have demonstrated the strength of Local Orientation Pyramid as a general representation for sample-guided progressive image coding. The discrimination performance and color constancy of local orientation have been successfully joined with the graceful degradation of histogram intersection. The number of steps to succeed (STS) was introduced as a performance figure for progressive image coding. It replaces the SNR which is the performance figure for traditional coders. STS for a face image was very low (7.5% of the total number of steps) and constant for a variety of photometric and geometric distortions. The additional compression ratio > 10 encourages the use of sample-guided progression for high-speed browsing of remote image databases.

References

[1] D. J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *Opt. Soc. Am. A*, 4:2379–2394, 1987.

[2] W. T. Freeman and E. H. Adelson. The design and use of steerable filters. *IEEE Trans. PAMI*, 13(9):891–906, 1991.

[3] B. V. Funt and G. D. Finlayson. Color constant color indexing. *IEEE Trans. PAMI*, 17(5):522–529, 1995.

[4] G. H. Granlund and H. Knutsson. *Signal Processing for Computer Vision*. Kluwer Academic Publishers, 1995.

[5] R. Jain, A. P. Pentland, and D. Petkovic. NSF–ARPA workshop on visual information mangement systems. Workshop report, Univ. of California at San Diego, 1995.

[6] E. H. Land. Recent advances in retinex theory. *Vision Res.*, 26:7–21, 1986.

[7] U. Mahlmeister, H. Pahl, and G. Sommer. Color–orientation indexing. In B. Jähne et al, editor, *18. DAGM*, pages 3–10. Springer, 1996.

[8] M. Michaelis and G. Sommer. Steerable filters in finite dimensional function spaces. Technical Report 9715, Christian–Albrechts–Universität zu Kiel, September 1997.

[9] A. Pentland, R.W. Picard, and S. Sclaroff. Photobook: Content–based manipulation of image databases. *Internat. Journ. of Computer Vision*, 18(3):233–254, 1996.

[10] B. Schiele and J. L. Crowley. Object recognition using multidimensional receptive field histograms. In *4. ECCV*, pages 610–619, April 1996.

[11] E. P. Simoncelli and W. T. Freeman. The steerable pyramid: a flexible architecture for multi–scale derivative computation. Technical report, GRASP Lab, Philadelphia, 1995.

[12] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heger. Shiftable multi–scale transforms. *IEEE Trans. Information Theory*, 38(2):587–607, 1992.

[13] M. J. Swain and D. H. Ballard. Color indexing. *Internat. Journ. of Computer Vision*, 7(1):11–32, 1991.