Кодування інформації

Дисципліна

Інформаційні технології

Лекція 2

Кодова таблиця ASCII
(American Standard Code for Information Interchange)

Dec	Hx	Oct	Char		Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr
0	0	000	NUL	(null)	32	20	040		Space	64	40	100	@	@	96	60	140	`	`
1	1	001	SOH	(start of heading)	33	21	041	!	!	65	41	101	A	A	97	61	141	a	a
2	2	002	STX	(start of text)	34	22	042	"	"	66	42	102	B	B	98	62	142	b	b
3	3	003	ETX	(end of text)	35	23	043	#	#	67	43	103	C	C	99	63	143	c	c
4	4	004	EOT	(end of transmission)	36	24	044	$	$	68	44	104	D	D	100	64	144	d	d
5	5	005	ENQ	(enquiry)	37	25	045	%	%	69	45	105	E	E	101	65	145	e	e
6	6	006	ACK	(acknowledge)	38	26	046	&	&	70	46	106	F	F	102	66	146	f	f
7	7	007	BEL	(bell)	39	27	047	'	'	71	47	107	G	G	103	67	147	g	g
8	8	010	BS	(backspace)	40	28	050	(	(	72	48	110	H	H	104	68	150	h	h
9	9	011	TAB	(horizontal tab)	41	29	051	)	)	73	49	111	I	I	105	69	151	i	i
10	A	012	LF	(NL line feed, new line)	42	2A	052	*	*	74	4A	112	J	J	106	6A	152	j	j
11	B	013	VT	(vertical tab)	43	2B	053	+	+	75	4B	113	K	K	107	6B	153	k	k
12	C	014	FF	(NP form feed, new page)	44	2C	054	,	,	76	4C	114	L	L	108	6C	154	l	l
13	D	015	CR	(carriage return)	45	2D	055	-	-	77	4D	115	M	M	109	6D	155	m	m
14	E	016	SO	(shift out)	46	2E	056	.	.	78	4E	116	N	N	110	6E	156	n	n
15	F	017	SI	(shift in)	47	2F	057	/	/	79	4F	117	O	O	111	6F	157	o	o
16	10	020	DLE	(data link escape)	48	30	060	0	0	80	50	120	P	P	112	70	160	p	p
17	11	021	DC1	(device control 1)	49	31	061	1	1	81	51	121	Q	Q	113	71	161	q	q
18	12	022	DC2	(device control 2)	50	32	062	2	2	82	52	122	R	R	114	72	162	r	r
19	13	023	DC3	(device control 3)	51	33	063	3	3	83	53	123	S	S	115	73	163	s	s
20	14	024	DC4	(device control 4)	52	34	064	4	4	84	54	124	T	T	116	74	164	t	t
21	15	025	NAK	(negative acknowledge)	53	35	065	5	5	85	55	125	U	U	117	75	165	u	u
22	16	026	SYN	(synchronous idle)	54	36	066	6	6	86	56	126	V	V	118	76	166	v	v
23	17	027	ETB	(end of trans. block)	55	37	067	7	7	87	57	127	W	W	119	77	167	w	w
24	18	030	CAN	(cancel)	56	38	070	8	8	88	58	130	X	X	120	78	170	x	x
25	19	031	EM	(end of medium)	57	39	071	9	9	89	59	131	Y	Y	121	79	171	y	y
26	1A	032	SUB	(substitute)	58	3A	072	:	:	90	5A	132	Z	Z	122	7A	172	z	z
27	1B	033	ESC	(escape)	59	3B	073	;	;	91	5B	133	[	[	123	7B	173	{	{
28	1C	034	FS	(file separator)	60	3C	074	<	<	92	5C	134	\	\	124	7C	174			|
29	1D	035	GS	(group separator)	61	3D	075	=	=	93	5D	135	]	]	125	7D	175	}	}
30	1E	036	RS	(record separator)	62	3E	076	>	>	94	5E	136	^	^	126	7E	176	~	~
31	1F	037	US	(unit separator)	63	3F	077	?	?	95	5F	137	_	_	127	7F	177	&#127;	DEL

Source: www.asciitable.com

Таблиці кодувань для української мови

(CP1125 стандартне кодування для DOS)

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
8	А 128	Б 129	В 130	Г 131	Д 132	Е 133	Ж 134	З 135	И 136	Й 137	К 138	Л 139	М 140	Н 141	О 142	П 143
9	Р 144	С 145	Т 146	У 147	Ф 148	Х 149	Ц 150	Ч 151	Ш 152	Щ 153	Ъ 154	Ы 155	Ь 156	Э 157	Ю 158	Я 159
A	а 160	б 161	в 162	г 163	д 164	е 165	ж 166	з 167	и 168	й 169	к 170	л 171	м 172	н 173	о 174	п 175
B	░ 176	▒ 177	▓ 178	│ 179	┤ 180	╡ 181	╢ 182	╖ 183	╕ 184	╣ 185	║ 186	╗ 187	╝ 188	╜ 189	╛ 190	┐ 191
C	└ 192	┴ 193	┬ 194	├ 195	─ 196	┼ 197	╞ 198	╟ 199	╚ 200	╔ 201	╩ 202	╦ 203	╠ 204	═ 205	╬ 206	╧ 207
D	╨ 208	╤ 209	╥ 210	╙ 211	╘ 212	╒ 213	╓ 214	╫ 215	╪ 216	┘ 217	┌ 218	█ 219	▄ 220	▌ 221	▐ 222	▀ 223
E	р 224	с 225	т 226	у 227	ф 228	х 229	ц 230	ч 231	ш 232	щ 233	ъ 234	ы 235	ь 236	э 237	ю 238	я 239
F	Ё 240	ё 241	Ґ 242	ґ 243	Є 244	є 245	І 246	і 247	Ї 248	ї 249	• 250	√ 251	№ 252	¤ 253	■ 254	nbsp 255

Таблиці кодувань для української мови

(KOI8-U стандарт для Інтернету – пошта і групи новин)

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
8	─ 128	│ 129	┌ 130	┐ 131	└ 132	┘ 133	├ 134	┤ 135	┬ 136	┴ 137	┼ 138	▀ 139	▄ 140	█ 141	▌ 142	▐ 143
9	░ 144	▒ 145	▓ 146	⌠ 147	■ 148	∙ 149	√ 150	≈ 151	≤ 152	≥ 153	nbsp 154	⌡ 155	° 156	² 157	· 158	÷ 159
A	═ 160	║ 161	╒ 162	ё 163	є 164	╔ 165	і 166	ї 167	╗ 168	╘ 169	╙ 170	╚ 171	╛ 172	ґ 173	╝ 174	╞ 175
B	╟ 176	╠ 177	╡ 178	Ё 179	Є 180	╢ 181	І 182	Ї 183	╦ 184	╧ 185	╨ 186	╩ 187	╤ 188	Ґ 189	╬ 190	© 191
C	ю 192	а 193	б 194	ц 195	д 196	е 197	ф 198	г 199	х 200	и 201	й 202	к 203	л 204	м 205	н 206	о 207
D	п 208	я 209	р 210	с 211	т 212	у 213	ж 214	в 215	ь 216	ы 217	з 218	ш 219	э 220	щ 221	ч 222	ъ 223
E	Ю 224	А 225	Б 226	Ц 227	Д 228	Е 229	Ф 230	Г 231	Х 232	И 233	Й 234	К 235	Л 236	М 237	Н 238	О 239
F	П 240	Я 241	Р 242	С 243	Т 244	У 245	Ж 246	В 247	Ь 248	Ы 249	З 250	Ш 251	Э 252	Щ 253	Ч 254	Ъ 255

Таблиці кодувань для української мови (CP1251 для ОС MS Windows)

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
8	Ђ 128	Ѓ 129	‚ 130	ѓ 131	„ 132	… 133	† 134	‡ 135	‌ 136	‰ 137	Љ 138	‹ 139	Њ 140	Ќ 141	Ћ 142	Џ 143
9	ђ 144	‘ 145	’ 146	“ 147	” 148	• 149	– 150	— 151	‌ 152	™ 153	љ 154	› 155	њ 156	ќ 157	ћ 158	џ 159
A	nbsp 160	Ў 161	ў 162	Ј 163	¤ 164	Ґ 165	¦ 166	§ 167	Ё 168	© 169	Є 170	« 171	¬ 172	shy 173	® 174	Ї 175
B	° 176	± 177	І 178	і 179	ґ 180	µ 181	¶ 182	· 183	ё 184	№ 185	є 186	» 187	ј 188	Ѕ 189	ѕ 190	ї 191
C	А 192	Б 193	В 194	Г 195	Д 196	Е 197	Ж 198	З 199	И 200	Й 201	К 202	Л 203	М 204	Н 205	О 206	П 207
D	Р 208	С 209	Т 210	У 211	Ф 212	Х 213	Ц 214	Ч 215	Ш 216	Щ 217	Ъ 218	Ы 219	Ь 220	Э 221	Ю 222	Я 223
E	а 224	б 225	в 226	г 227	д 228	е 229	ж 230	з 231	и 232	й 233	к 234	л 235	м 236	н 237	о 238	п 239
F	р 240	с 241	т 242	у 243	ф 244	х 245	ц 246	ч 247	ш 248	щ 249	ъ 250	ы 251	ь 252	э 253	ю 254	я 255

Юнікод

- Кожному юнікод-символу поставлено у відповідність невід'ємне ціле число, що називається його кодової позицією (code point).
- Кодовий простір Юникода складається из 1 114 112 кодових позицій в діапазоні від 0 до 10FFFF. З них до дев'ятої версії стандарту значення присвоєно лише 128 237. Частина простору зарезервована для приватного використання і консорціум Юникода обіцяє ніколи не привласнювати значення позицій з цих спеціальний областей.
- U+0020 (пробіл)
- U+0300 и U+0301 (наголоси сильний і слабкий)
- U+E0001, U+E0020–U+E007E, и U+E007F (мовні теги, американський і британський варіанти англійської)

Завадостійке кодування

- Можливість появи збоїв при роботі електронної техніки (електромагнітні завади, малий рівень сигналу, старіння радіоелементів, радіоактивне випромінювання).
- Код, призначений для знаходження одної помилки (інверсії біта). Додається контрольний біт. Контрольний біт має бути таким, щоб сумарна кількість одиниць у машинному слові була парною.
- 10111100 101111001

$$k = \begin{cases} 1, & b_1 \oplus b_2 \oplus \ldots \oplus b_n = 1 \\ 0, & b_1 \oplus b_2 \oplus \ldots \oplus b_n = 0 \end{cases}$$

Код Хеммінга

n – значний код Хеммінга має m - інформаційних розрядів і k – контрольних. Кількість контрольних розрядів повинно задовольняти умові

$$k \geq \log_2(k + 1)$$

тому

$$k \leq n - \log_2(k + 1)$$

Код Хеммінга будується наступним чином: до інформаційних розрядів кодової комбінації додається обрахована за формулою потрібна кількість контрольних розрядів, які формуються шляхом підрахунку парності суми одиниць для певних груп інформаційних розрядів. При отриманні такої кодової комбінації з інформаційних та контрольних розрядів шляхом аналогічних підрахунків парності складають коректуюче число, яке дорівнює нулю при відсутності помилки, або вказує номер помилкового розряду.

Процес кодування за Хеммінгом

Розглянемо процес кодування. Нехай перший контрольний розряд має непарний порядковий номер, визначимо його при кодуванні таким чином, щоб сума одиниць всіх розрядів з непарними порядковими номерами дорівнювала нулю. Така операція може бути записана у вигляді наступного співвідношення:

$$a_1 = a_1 \oplus a_3 \oplus a_5 \oplus a_7 \ldots = 0$$

де a_1, a_3, a_5, a_7 – двійкові символи, розташовані в розрядах з номерами 1,3,5,7,…

Поява одиниці у другому розряді (справа) коректую чого числа означає помилку у тих розрядах кодової комбінації, порядкові номера яких (2,3,6,7,…) у двійковому зображені мають одиницю у другому справа розряді. Тому друга операція кодування, яка дозволяє знайти другій контрольний розряд, має вигляд

$$a_2 = a_2 \oplus a_3 \oplus a_6 \oplus a_7 \ldots = 0$$

Аналогічно

$$a_3 = a_4 \oplus a_5 \oplus a_6 \oplus a_7 \ldots = 0$$

$$a_4 = a_8 \oplus a_9 \oplus a_{10} \oplus a_{11} \ldots = 0$$

Декодування за Хеммінгом

Після прийому кодової комбінації спільно зі сформованими контрольними розрядами виконуються ті ж операції підрахунку, що було описано вище, а отримане число

$$E_k E_{k-1} \ldots E_2 E_1$$

Вважається коректуючим, причому при $E_k E_{k-1} \ldots E_2 E_1 = 0$ помилки відсутні, а при наявності помилок нерівними виявляються ті суми E_i, в утворенні яких приймав участь помилковий розряд. Коректуюче число при цьому дорівнює порядковому номеру цього розряду.

Вибір місця для контрольних розрядів у кожній з кодових комбінацій визначається таким чином, щоб контрольні розряди приймали участь тільки в одній операції підрахунку парності. Такими позиціями є цілі степені двійки: 1,2,4,8,16,…

Приклад кодування за Хеммінгом

Приклад. Складемо шестизначний код Хеммінга для n=6, $k>log_2 7$, k=3, $m=n-k$=3

Цифра	Двійковий код цифри	Код Хеммінга 654321
0	000	000000
1	001	000111
2	010	011001
3	011	011110
4	100	101010
5	101	101101
6	110	110011
7	111	110100

QR-коди

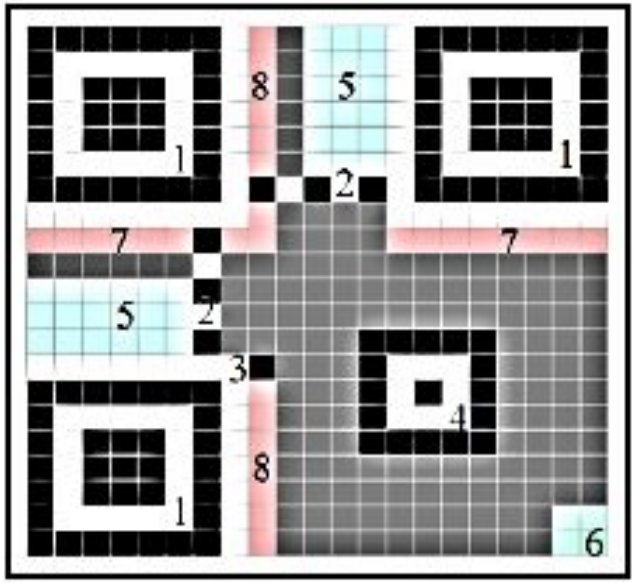
QR-код (англ. Quick response - швидкий відгук) - матричний код, розроблений японською компанією «Denso-Wave» в 1994 році. Місткість однієї кодової матриці не є безмежною. Максимальна кількість символів, які можна помістити в одну матрицю, залежить від виду інформації, що кодується (цифри, букви, двійковий код, ієрогліфи), рівня корекції помилок і версії QR-коду. Для версії 40 в одній матриці при мінімальному рівні корекції помилок можна розмістити 7089 десяткових цифр, або 4296 букв і цифр, або 2953 двійкових даних, або 1817 ієрогліфів.

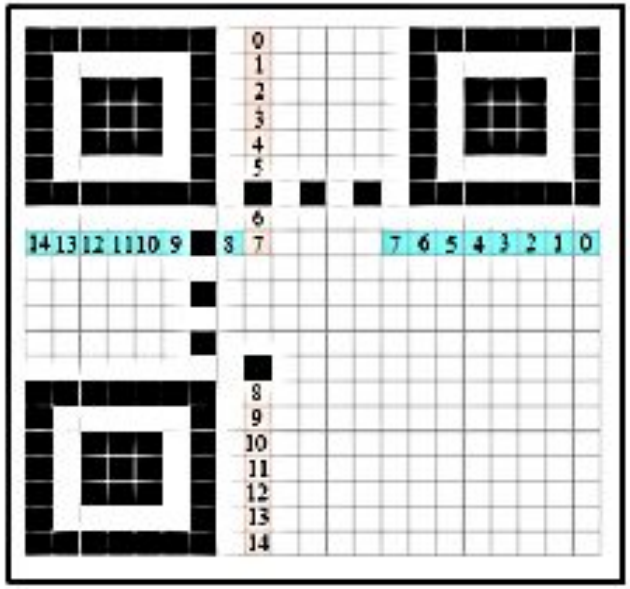
Даний код має високу стійкість по відношенню до спотворень. Навіть при наявності пошкоджень зображення на площі 30% все ще є можливість безпомилково зчитувати інформацію.

Для того щоб зробити туристичні маршрути більш інформативними, на вулицях Львова розміщено понад 80 зображень QR-кодів. Ними обладнані комерційні та культурні об'єкти, в тому числі пам'ятники, архітектурні споруди.

QR-код не є єдиним варіантом двомірного штрих-коду. Відомі й інші формати: DataMatrix, ScanLife EZcode і Microsoft Tag (Tag).

Порядок розміщення елементів матриці QR-кода



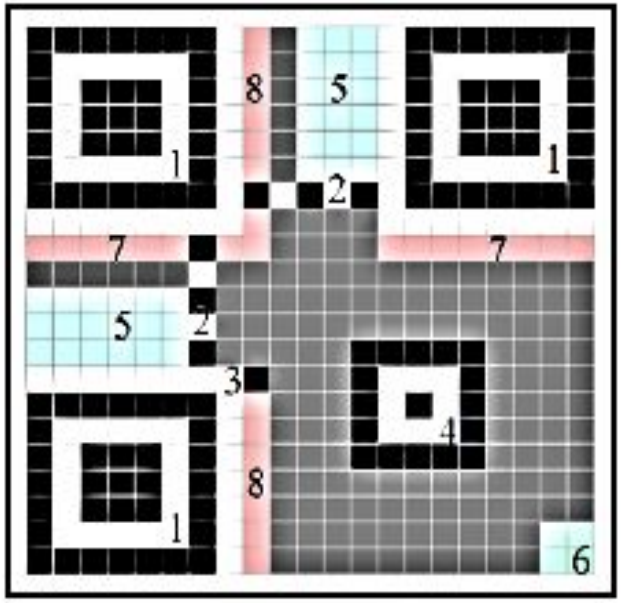


Три квадрати в кутах зображення QR-коду дозволяють пристрою правильно визначати його просторову орієнтацію. Чорні та білі пікселі зображення перетворюються зчитувальним пристроєм у двійкові числа. Потім проводиться обробка цифрової інформації. На матриці розміщується системна інформація, інформаційні біти, коригувальні біти. При обробці зображення використовується алгоритм Ріда-Соломона, який дозволяє усунути третину можливих спотворень матриці.

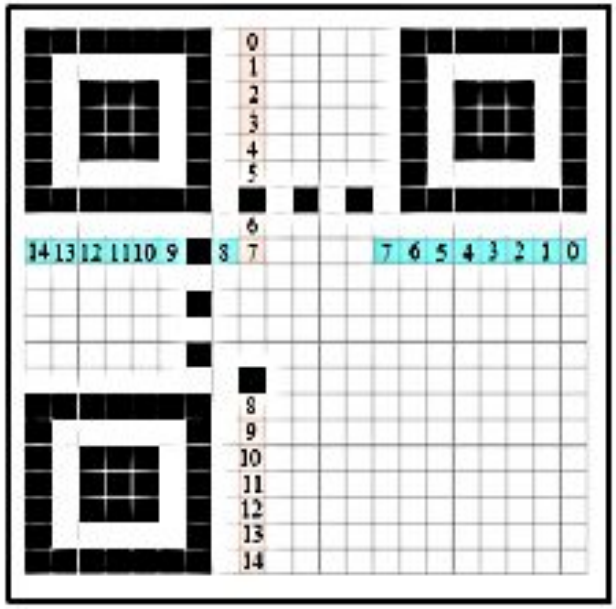
На кожній матриці є три покажчика орієнтації 1 (ПО). Між трьома ПО проходять дві лінії синхронізації 2, які складаються з білих і чорних пікселів, що чергуються. Поблизу лівого нижнього покажчика орієнтації розміщується чорний піксель 3.

На сьогодні створено 40 версій QR-кодів. На матрицях QR-кодів, починаючи з версії 2, є мітки центрування 4. У версії 1 такої мітки немає, але в старших версіях QR-коду можна нарахувати до 46 таких міток.

Порядок розміщення елементів матриці QR-кода



В областях 5 для версій старше шостої розміщується інформація про версії коду. У чотирьох пікселях 6 розташовано код, що визначає, який вид даних міститься в матриці. Системна інформація розміщується в областях 7 і 8.



Інформація про рівень корекції помилок і вигляду використаної маски розташовується в системному рядку. Розряди 14 і 13 містять відомості про використаний рівень корекції помилок. Розряди 12, 11 і 10 показують, яка маска використана для накладання на інформаційні та коригувальні біти. Три біти дозволяють сформувати коди для восьми різних масок. Системна інформація дублюється. Один раз 15 системних бітів розміщують в стовпці 8 (рахунок елементів ведеться, починаючи з 0 з лівого верхнього кута матриці). Другий раз системну інформацію розміщують в рядку 7.

Рівні корекції QR-кодів

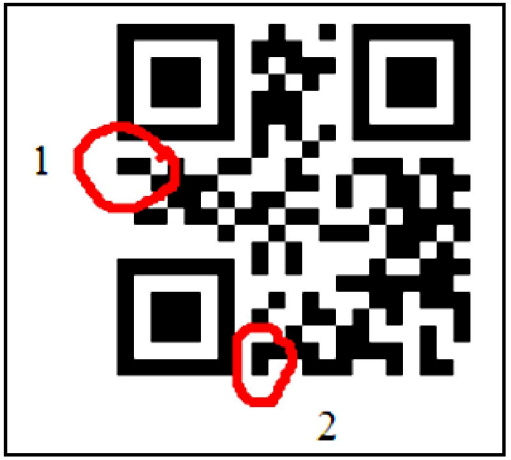
Рівень корекції помилок	Код	Припустимі спотворення
L	01	7%
M	00	15%
Q	11	25%
H	10	30%

Рівні L і M рекомендується використовувати для друку кодів на запрошеннях, білетах, оголошеннях, візитках, плакатах, рекламних щитах і т.д. Рівні Q і H застосовують для маркування промислових деталей, коли існує висока ймовірність пошкодження зображення коду.

7%		15%	
25%		30%	

Проба

Визначення рівня корекції



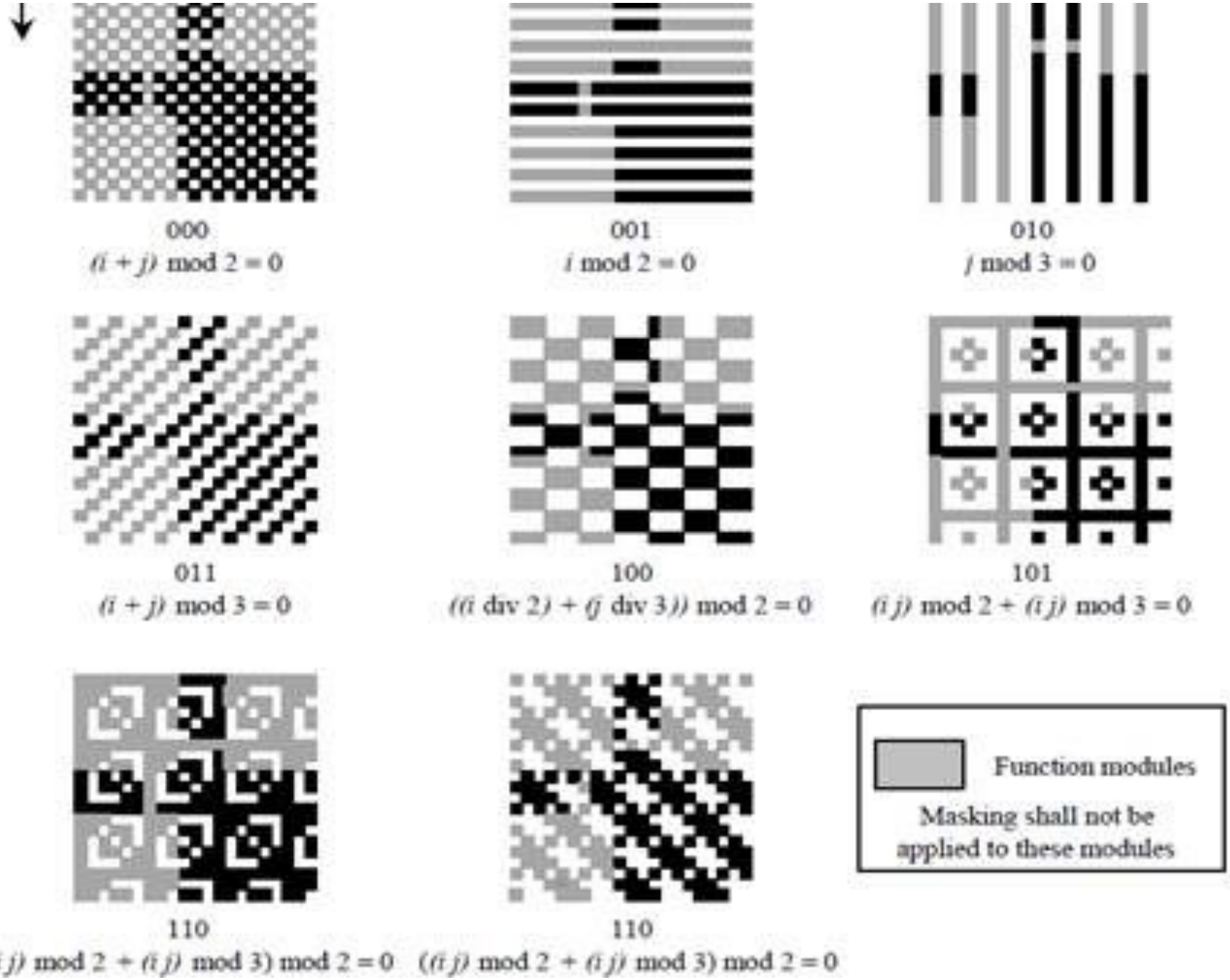
Так як системна інформація дублюється, то відомості про наявну коректуючу здатність коду можна відшукати в двох місцях матриці. Один раз 15 біт системної інформації розміщено в рядку 1 (відлік зліва направо), а другий раз - в стовпці 2 (відлік знизу-вгору). Природно, що ці два значення однакові. Відомості про рівень корекції помилок містяться в перших двох бітах системної інформації. Ці два біта розташовані в 14 і 13 розрядах системного рядка.

Щоб визначити рівень корекції, потрібно ці два біта додати за правилом «Виключного Або» (XOR) з двійковою маскою 10. В результаті додавання отримаємо: 11 $\oplus$10 =01. Код 01 відповідає рівню корекції помилок L (7%).

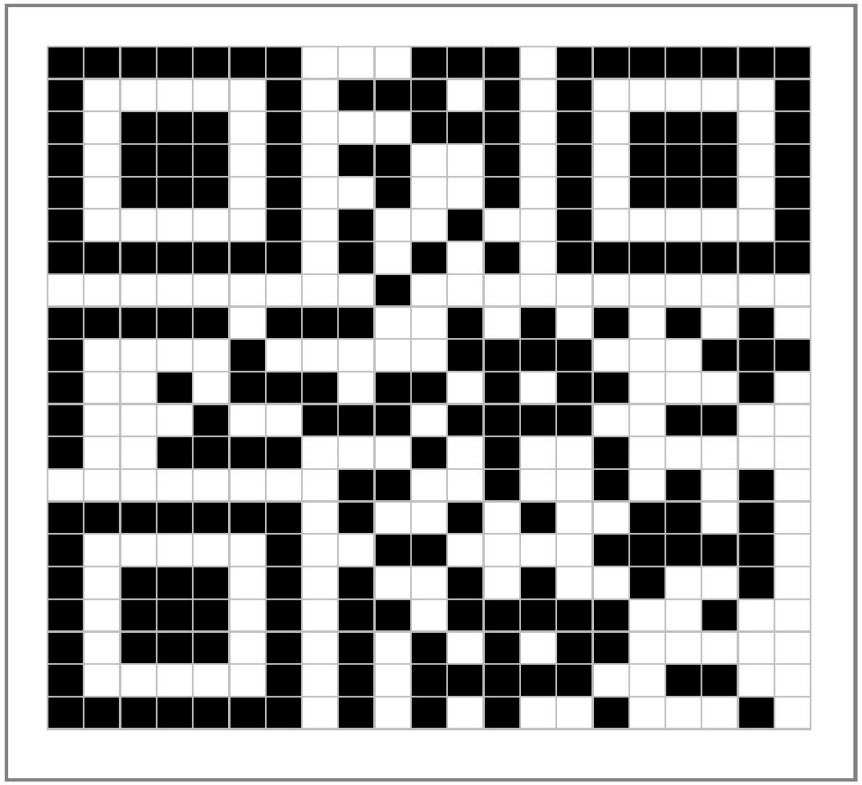
Маски QR - кодів

Для підвищення надійності сканування матриці додатково перетворюють. Метою цього перетворення є формування такого рисунку (конфігурації) матриці, на якому було б якомога менше суміжних пікселів однакового кольору. Мета досягається наступним чином: формують 8 матриць з однаковим інформаційним змістом, але різного виду (різної форми). Для цього використовують 8 заздалегідь підготовлених масок, які підсумовують з вихідною матрицею за правилом XOR. Отримані матриці почергово перетворюють, в результаті вибирають одну матрицю. Відбір матриць здійснюють шляхом обчислення штрафних балів. Для оптимальної (найбільш «строкатої») матриці число штрафних балів буде мінімальним серед восьми чисел. Вісім масок кодуються трирозрядними числами від 000 до 111. Математично маски описуються формулами, наведеними в таблиці.

000	$(i+j) \bmod 2 \equiv 0$
001	$i \bmod 2 \equiv 0$
010	$j \bmod 3 \equiv 0$
011	$(i+j) \bmod 3 \equiv 0$
100	$((i \ div\ 2)+(j \ div\ 3)) \bmod 2 \equiv 0$
101	$(i \cdot j) \bmod 2+(i \cdot j) \bmod 3 \equiv 0$
110	$((i \cdot j) \bmod 2+(i \cdot j) \bmod 3) \bmod 2 \equiv 0$
111	$((i \cdot j) \bmod 3+(i \cdot j) \bmod 2) \bmod 2 \equiv 0$


000
(i + j) mod 2 = 0
001
i mod 2 = 0
010
j mod 3 = 0
011
(i + j) mod 3 = 0
100
((i div 2) + (j div 3)) mod 2 = 0
101
(i j) mod 2 + (i j) mod 3 = 0
110
((i j) mod 2 + (i j) mod 3) mod 2 = 0
110
((i j) mod 2 + (i j) mod 3) mod 2 = 0
Function modules
Masking shall not be applied to these modules

Декодування QR - кода



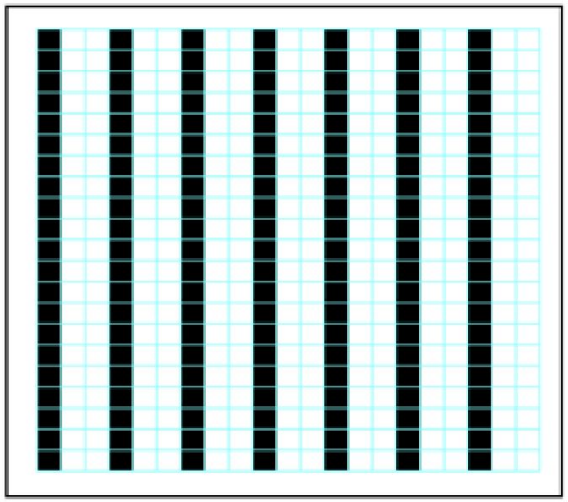
Спочатку потрібно визначити, який вид маски використаний при її кодуванні. У розрядах 12, 11 і 10 системного рядка записано бінарне слово 111. Додавання зі статичною маскою 101 дає двійкове число 010. По таблиці кодів маски легко визначити формулу, використану для формування маски: **jmod 3 ≡0**.

За цією формулою можна визначити вид розрахункової маски. Цей рисунок потрібно трактувати наступним чином: якщо на масці є чорний піксель, то піксель матриці з такими ж координатами потрібно проінвертувати.

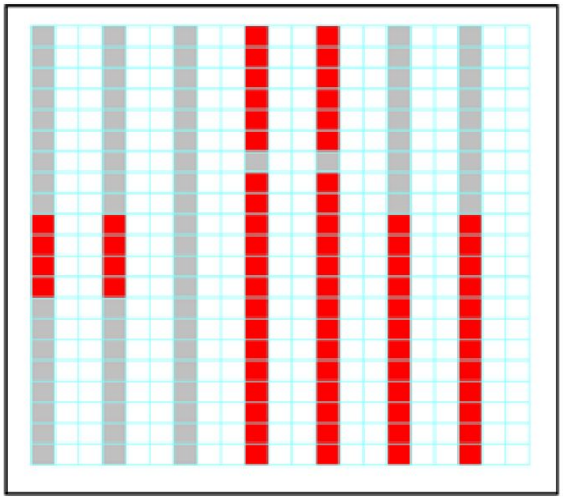
У цього правила велике число винятків. На рис. показані сірим кольором модулі, які інвертувати не слід.

Від зміни захищені області матриці, де розташовані покажчики орієнтації, лінії синхронізації, системна інформація (сірі модулі). Скориговану таким чином розрахункову маску назвемо фактичної маскою.

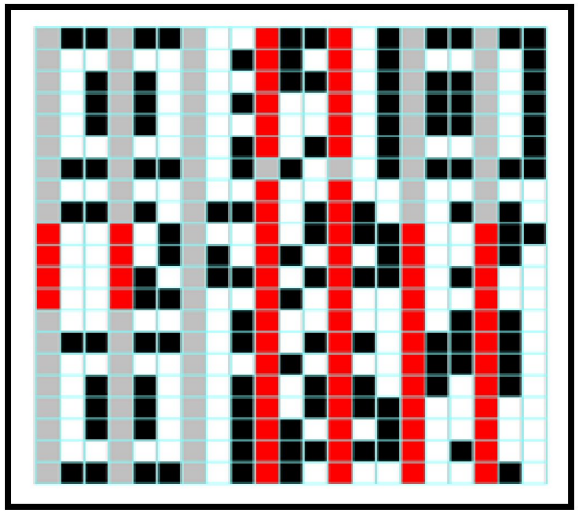
Накладання маски на матрицю



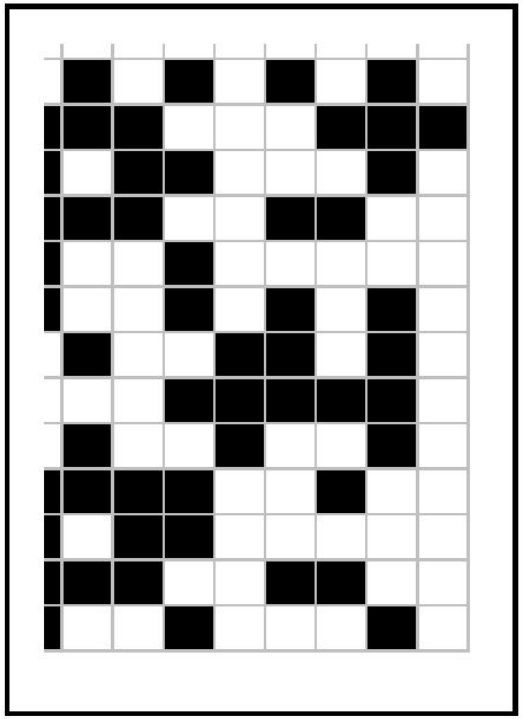
а)

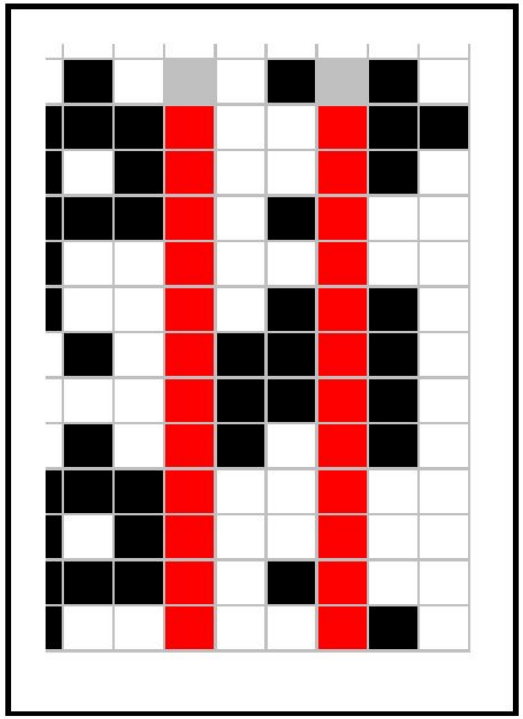


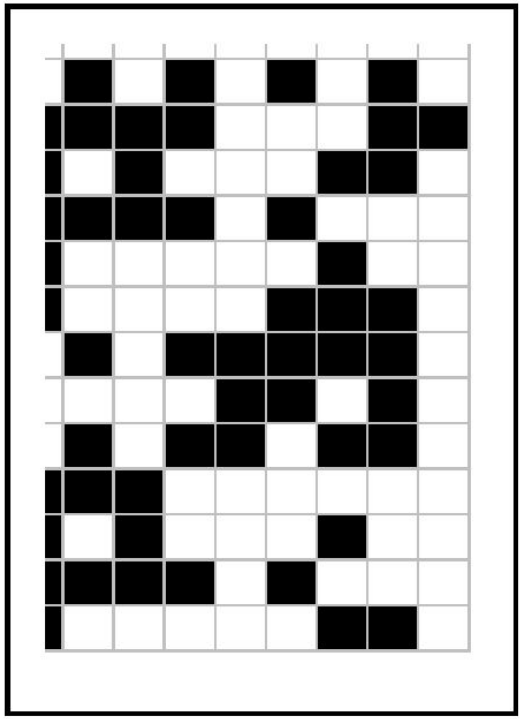
в)



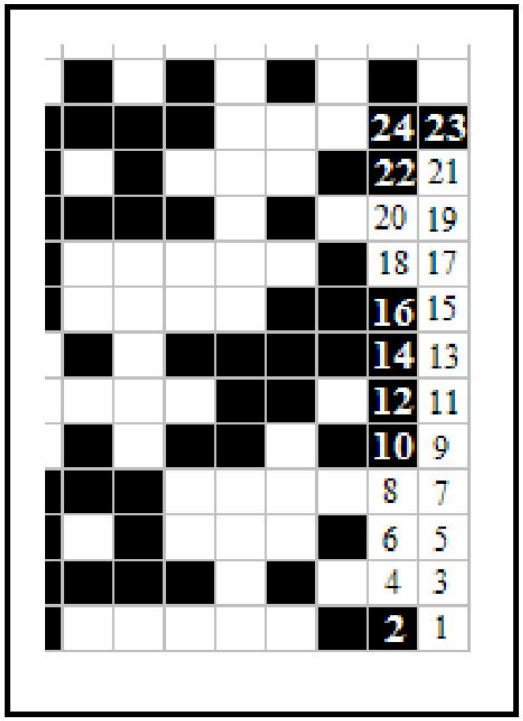
Результат накладання маски на матрицю

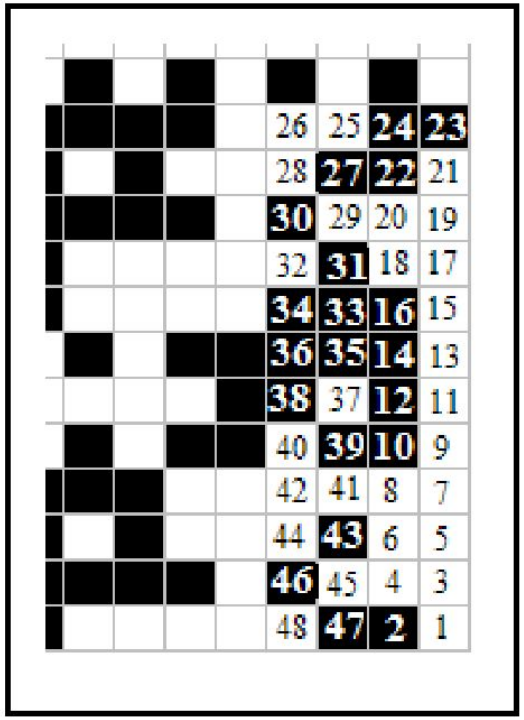


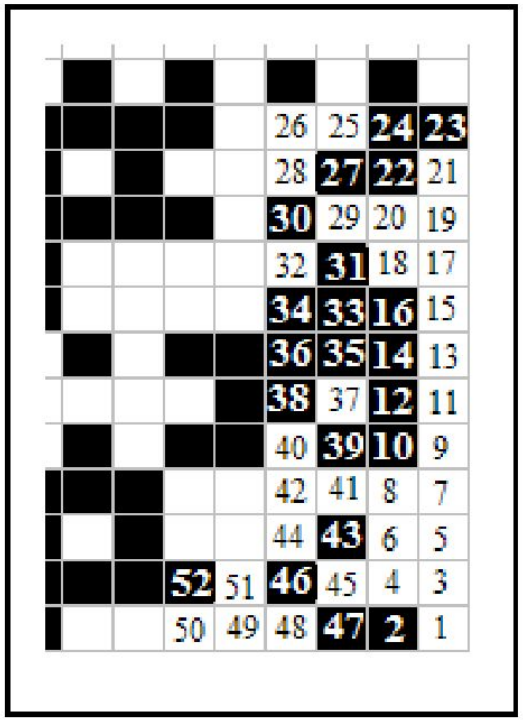




Ручне декодування матриці







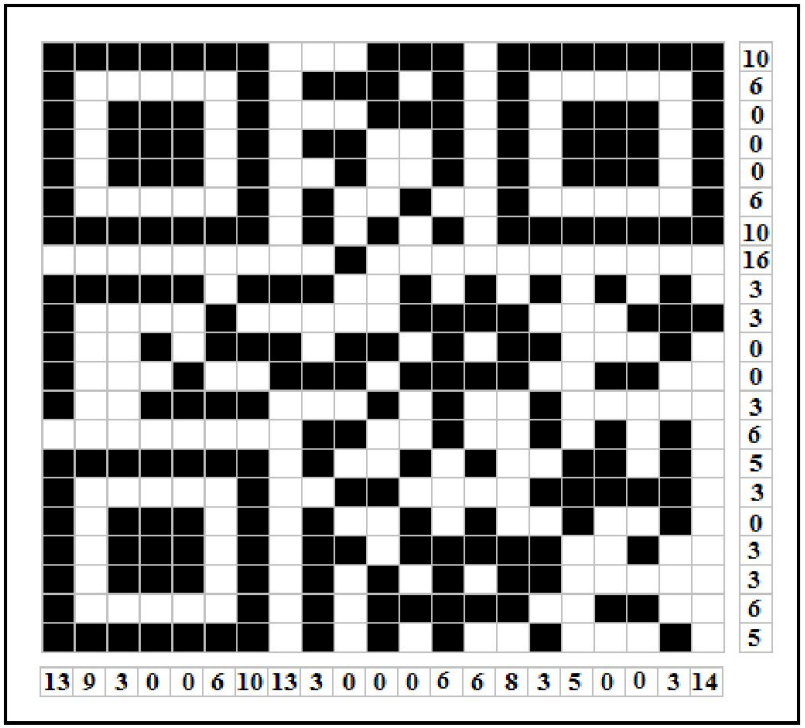
Чотири біта в клітинках 1 ... 4 дають код 0100, який говорить про те, що в даній матриці використовується двійковий формат даних.

Відповідно до двійкового формату даних наступні 8 біт (клітинки 5 ... 12) вказують, скільки символів міститься в повідомленні. Двійкове число 00000101 говорить про те, що дане повідомлення містить п'ять символів. У наступних восьми клітинках 13 ... 20 (див. рисунок 11 а)) розміщено двійкове слово 01010000. Запис цього байта в десятковій системі числення дає число 80. За таблицею CP-1251 легко визначити, що цьому коду відповідає латинська буква

Результат декодування

Номери пікселів	Біти	Десяткове число	Символ
13…20	01010000	80	P
21…28	01110000	114	r
29…36	01101111	111	o
37…44	01100010	98	b
45…52	01100001	97	a

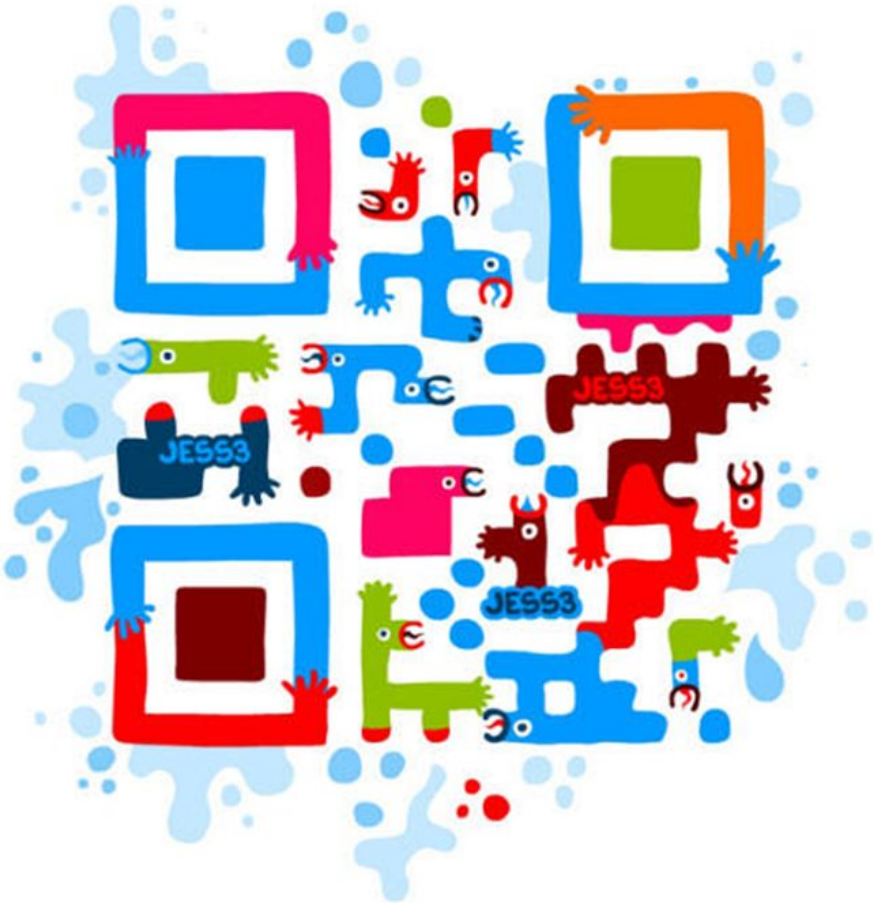
Правила кодування (розрахунок штрафів)



При кодуванні повідомлення формують вісім матриць з однаковим змістом, але різної форми. Для вибору однієї з восьми конкуруючих матриць в стандарті визначено чотири правила розрахунку штрафів. Відповідно до першого правила штраф призначається за кожну групу з п'яти або більше одноколірних пікселів в одному рядку. Штрафні функції по стовпцях обчислюються аналогічно. Друге правило дає штраф за кожен одноколірний квадрат розміром 2x2 пікселя. Третє правило визначає штраф, якщо на матриці є області, які схожі на покажчики визначення орієнтації. Четверте правило нараховує штраф, якщо більше половини пікселів одного кольору. Метою розрахунку штрафних балів є вибір найбільш «строкатої» матриці з восьми можливих варіантів.

Для розглянутого прикладу сумарна кількість штрафних балів за рядками дорівнює 88, а за стовпцями 102. Загальна кількість штрафних балів, що обчислено за першим правилом для даної маски, дає число 190.

Використовуючи всі чотири правила розрахунку штрафних балів, знаходять сумарну кількісну оцінку даної маски. Подібні оцінки отримують для всіх восьми масок. Серед цих оцінок вибирають мінімальну оцінку і для остаточного кодування використовують маску, яка має мінімальне число штрафних балів.



BS ISO/IEC 18004:2006. Information technology. Automatic identification and data capture techniques. QR Code 2005 bar code symbology specification. London: BSI. 2007. p. 126. ISBN 978-0-580-67368-9.

Боб написав повідомлення для Аліси, використавши для цього таблицю CP1251, тобто кожна літера була записана за допомогою 1 байта.
Далі Боб застосував завадостійке кодування за Хеммінгом і надіслав повідомлення.
Єва не хотіла, щоб Аліса отримала правильне повідомлення і внесла у код кожної літери по одній похибці, інвертувавши по одному біту.
Аліса отримала наступний код:

110011011001 перша літера (декодують студенти N mod5 = 1)

111011000110 друга літера (декодують студенти N mod5 = 2)

101010001110 третя літера (декодують студенти N mod5 = 3)

111010101100 четверта літера (декодують студенти N mod5 = 4)

111100101000 п’ята літера (декодують студенти N mod5 = 0)

Яке повідомлення надіслав Боб? Кожну букву записати у вигляді десяткового числа.