Отчет по второму практическому заданию курса Байесовские методы в машинном обучении

студентки 4 курса ВШЭ
Замышевской Арины Александровны

1 Описание модели

Дана выборка $X = \{X_k\}_{k=1}^K$ сильно зашумленных черно-белых изображений размера $H \times W$ пикселей. Каждое из этих изображений содержит один и тот же неподвижный фон и лицо преступника в неизвестных координатах, при этом лицо попадает в любое изображение целиком. Будем считать, что изображение лица имеет прямоугольную форму размера $h \times w$ пикселей.

Введем следующие обозначения:

- $X_k(i, j)$ — пиксель k-ого изображения;
- $B \in \mathbb{R}^{H \times W}$ — изображение чистого фона без лица преступника, $B(i, j)$ — пиксель этого изображения;
- $F \in \mathbb{R}^{h \times w}$ — изображение лица преступника, $F(i, j)$ — пиксель этого изображения;
- $d_k = (d_{h_k}, d_{w_k})$ — координаты верхнего левого угла изображения лица на k-ом изображении (d_{h_k} по вертикали, d_{w_k} по горизонтали), $d = (d_1, \ldots, d_K)$ — набор координат для всех изображений выборки.

Также будем считать, что шум на изображении независим для каждого пикселя и принадлежит нормальному распределению $N(0, s^2)$, где s — стандартное отклонение. Таким образом, для одного изображения имеем:

$$p(X_k|d_k, \theta) = \prod_{i,j} \begin{cases} N(X_k(i, j)|F(i - d_{h_k}, j - d_{w_k}), s^2), & \text{если } (i, j) \in \text{faceArea}(d_k) \\ N(X_k(i, j)|B(i, j), s^2), & \text{иначе} \end{cases}$$

где $\theta = \{B, F, s^2\}$, а $\text{faceArea}(d_k) = \{(i, j)|d_{h_k} \le i \le d_{h_k} + h - 1, d_{w_k} \le j \le d_{w_k} + w - 1\}$.

Распределение на неизвестные координаты лица на изображении зададим общим для всех изображений с помощью матрицы параметров $A \in \mathbb{R}^{H-h+1, W-w+1}$ следующим образом:

$$p(d_k|A) = A(d_{h_k}, d_{w_k}), \quad \sum_{i,j} A(i, j) = 1,$$

где $A(i, j)$ — элемент матрицы A. В итоге имеем следующую совместную вероятностную модель:

$$p(X, d|\theta, A) = \prod_k p(X_k|d_k, \theta) p(d_k|A).$$

2 Теория

Тут просят вывести формулы для подсчета следующих величин:

2.1 Апостериорное распределение на координаты лица на изображениях

На E-шаге вычисляется оценка на апостериорное распределение на координаты лица на изображениях:

$$q(d) = p(d|X, \theta, A) = \prod_k p(d_k|X_k, \theta, A).$$

По формуле Байеса можем записать:

$$p(d_k|X_k, \theta, A) = \frac{p(X_k|d_k, \theta)p(d_k|A)}{p(X_k|\theta, A)}.$$

Скорее для себя расписываю, что тут $p(X_k|d_k, \theta)$ — правдоподобие X_k при условии d_k и параметров θ; $p(d_k|A)$ — априорное распределение координат d_k с параметрами A; $p(X_k|\theta, A)$ — нормализующая штука, вот она ниже:

$$p(X_k|\theta, A) = \sum_{d_k} p(X_k|d_k, \theta)p(d_k|A).$$

Итог:

$$q(d_k) = p(d_k|X_k, \theta, A) = \frac{p(X_k|d_k, \theta)p(d_k|A)}{\sum_{d'_k} p(X_k|d'_k, \theta)p(d'_k|A)}.$$

2.2 Точечные оценки на параметры

На M-шаге вычисляются точечные оценки на параметры A, $\theta = \{F, B, s^2\}$:

1. Для оценки параметров A необходимо максимизировать:

$$E_{q(d)} \left[\log p(X, d|\theta, A)\right] \to \max A.$$

$$E_{q(d)} \left[\log p(X, d|\theta, A)\right] = E_{q(d)} \left[\log p(X|d, \theta)\right] + E_{q(d)} \left[\log p(d|A)\right].$$

Теперь подставим выражение для правдоподобия:

$$E_{q(d)} \left[\log p(X|d, \theta)\right] = \sum_{k=1}^{K} E_{q(d_k)} \left[\log p(X_k|d_k, \theta)\right].$$

$$p(d_k|A) = A(d_{h_k}, d_{w_k})$$

$$E_{q(d)} \left[\log p(d|A)\right] = \sum_{k=1}^{K} E_{q(d_k)} \left[\log A(d_{h_k}, d_{w_k})\right].$$

Дальше производная и приравнение к нулю для максимизации:

$$\frac{\partial}{\partial A} E_{q(d)} \left[\log p(X, d|\theta, A)\right] = 0.$$

Откуда

$$A(i, j) = \frac{\sum_{k=1}^{K} q_k(d_k) \cdot I(d_k = (i, j))}{\sum_{k=1}^{K} q_k(d_k)},$$

тут $I(d_k = (i, j))$ — функция, которая $= 1$, если координаты d_k совпадают с (i, j).

И вот итог:

$$A(i, j) = \frac{1}{K} \sum_{k=1}^{K} q_k(d_k) \cdot I(d_k = (i, j)).$$

2. Для оценки изображения лица F:

$$E_{q(d)}\left[\log p(X, d|\theta, A)\right] \to \max F.$$

$$\log p(X, d|\theta, A, F) = \sum_{k=1}^{K} \left(E_{q(d_k)}\left[\log p(X_k|d_k, \theta, A, F)\right] + E_{q(d_k)}\left[\log p(d_k|A)\right]\right).$$

Правдоподобие для X_k:

$$p(X_k|d_k, \theta, A, F) = \prod_{(i,j)} \mathcal{N}(X_k(i,j)|F(i-d_{h_k}, j-d_{w_k}), s^2) + \mathcal{N}(X_k(i,j)|B(i,j), s^2),$$

$$E_{q(d)}\left[\log p(X, d|\theta, A, F)\right] =$$

$$= \sum_{k=1}^{K} E_{q(d_k)} \left[\sum_{(i,j)} \log\left(\mathcal{N}(X_k(i,j)|F(i-d_{h_k}, j-d_{w_k}), s^2) + \mathcal{N}(X_k(i,j)|B(i,j), s^2)\right)\right].$$

$$\frac{\partial}{\partial F} E_{q(d)}\left[\log p(X, d|\theta, A, F)\right] = 0.$$

$$-\frac{1}{s^2}\left(F - \sum_{k=1}^{K} E_{q(d_k)}\left[X_k(i,j) \cdot \delta(d_k)\right]\right) = 0,$$

$\delta(d_k)$ — индикаторная функция, она равна 1, если пиксель принадлежит лицу.

$$F = \frac{1}{K}\sum_{k=1}^{K} E_{q(d_k)}\left[X_k \cdot \delta(d_k)\right].$$

3. Для оценки изображения фона B:

$$E_{q(d)}\left[\log p(X, d|\theta, A)\right] \to \max B.$$

Поехали:

$$E_{q(d)}\left[\log p(X, d|\theta, A)\right] = \sum_{k=1}^{K} E_{q(d_k)}\left[\log p(X_k|d_k, \theta)\right].$$

$$E_{q(d)}\left[\log p(X, d|\theta, A)\right] = -\frac{KHW}{2}\log(2\pi s^2) -$$

$$-\frac{1}{2s^2}\sum_{k=1}^{K} E_{q(d_k)}\left[\sum_{(i,j)\in \text{face}(d_k)} (X_k(i,j) - F(i-d_{h_k}, j-d_{w_k}))^2 + \sum_{(i,j)\notin \text{face}(d_k)} (X_k(i,j) - B(i,j))^2\right].$$

Снова производная и ноль...

$$\frac{\partial}{\partial B} E_{q(d)}\left[\log p(X, d|\theta, A)\right] = 0.$$

$$B(i,j) = \frac{\sum_{k=1}^{K}\sum_{(i,j)\notin \text{face}(d_k)} q_k(d_k) X_k(i,j)}{\sum_{k=1}^{K}\sum_{(i,j)\notin \text{face}(d_k)} q_k(d_k)}.$$

И вот итог:

$$B = \frac{1}{K}\sum_{k=1}^{K} X_k \cdot (1 - q_k(d_k)),$$

тут q_k — свертка с индикаторной функцией, функция теперь исключает пиксели, принадлежащие лицу.

4. Для оценки стандартного отклонения s^2:

$$E_{q(d)}\left[\log p(X,d|\theta,A)\right] \to \max s^2.$$

$$p(X_k|d_k,\theta) = \prod_{(i,j)\in \text{face}(d_k)} \mathcal{N}(X_k(i,j)|F(i-d_{h_k},j-d_{w_k}),s^2) \prod_{(i,j)\notin \text{face}(d_k)} \mathcal{N}(X_k(i,j)|B(i,j),s^2).$$

$$E_{q(d)}\left[\log p(X,d|\theta,A)\right] = \sum_{k=1}^{K} E_{q(d_k)}\left[\log p(X_k|d_k,\theta)\right].$$

Делаем подстановку снова:

$$E_{q(d)}\left[\log p(X,d|\theta,A)\right] = -\frac{KHW}{2}\log(2\pi s^2) -$$

$$-\frac{1}{2s^2}\sum_{k=1}^{K} E_{q(d_k)}\left[\sum_{(i,j)\in \text{face}(d_k)} (X_k(i,j)-F(i-d_{h_k},j-d_{w_k}))^2 + \sum_{(i,j)\notin \text{face}(d_k)} (X_k(i,j)-B(i,j))^2\right].$$

Снова производная и приравнение к нулю, все стандартно:

$$\frac{\partial}{\partial s^2} E_{q(d)}\left[\log p(X,d|\theta,A)\right] = -\frac{KHW}{2s^2} +$$

$$+\frac{1}{2(s^2)^2}\sum_{k=1}^{K} E_{q(d_k)}\left[\sum_{(i,j)\in \text{face}(d_k)} (X_k(i,j)-F(i-d_{h_k},j-d_{w_k}))^2 + \sum_{(i,j)\notin \text{face}(d_k)} (X_k(i,j)-B(i,j))^2\right] = 0.$$

$$s^2 = \frac{1}{KHW}\sum_{k=1}^{K} E_{q(d_k)}\left[\sum_{(i,j)\in \text{face}(d_k)} (X_k(i,j)-F(i-d_{h_k},j-d_{w_k}))^2 + \sum_{(i,j)\notin \text{face}(d_k)} (X_k(i,j)-B(i,j))^2\right].$$

2.3 Нижняя оценка на логарифм неполного правдоподобия

Нижняя оценка на логарифм неполного правдоподобия такая:

$$L(q,\theta,A) = E_{q(d)}\left[\log p(X,d|\theta,A)\right] - E_{q(d)}\left[\log q(d)\right] \to \max q,\theta,A.$$

Тут $E_{q(d)}\left[\log p(X,d|\theta,A)\right]$ — ожидаемое значение логарифма полного правдоподобия, его вот мы и хотим максимизировать, а $E_{q(d)}\left[\log q(d)\right]$ — энтропия апостериорного распределения $q(d)$, которая штрафует за сложность.

По формуле Байеса полное правдоподобие:

$$p(X,d|\theta,A) = p(X|d,\theta)p(d|A).$$

Отсюда и нижняя оценка:

$$L(q,\theta,A) = E_{q(d)}\left[\sum_{k=1}^{K}\log p(X_k|d_k,\theta) + \sum_{k=1}^{K}\log p(d_k|A)\right] - E_{q(d)}\left[\log q(d)\right].$$

3 Анализ

3.1 Влияние начального приближения на результаты работы

Протестируйте полученный EM алгоритм на сгенерированных данных. Оцените сильно ли влияет начальное приближение на параметры на результаты работы.

Стоит ли для данной задачи запускать EM алгоритм из разных начальных приближений?

Вот так выглядят мои сгенерированные данные: это фон + лицо (смайлик в очках), наложенное в рандомном месте на фон.

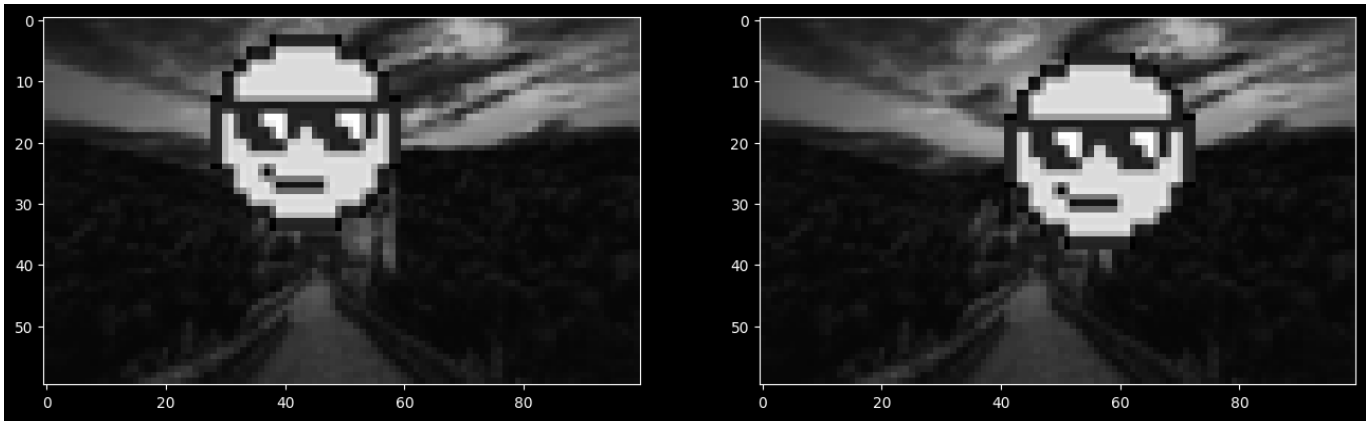

Рис. 1: Два примера сгенерированных данных

Дальше я запустила EM алгоритм с разными начальными приближениями 10 раз (можно и больше, но тогда график сложнее читать).

По итогам работы получила вот такие картинки лица (смайлика):

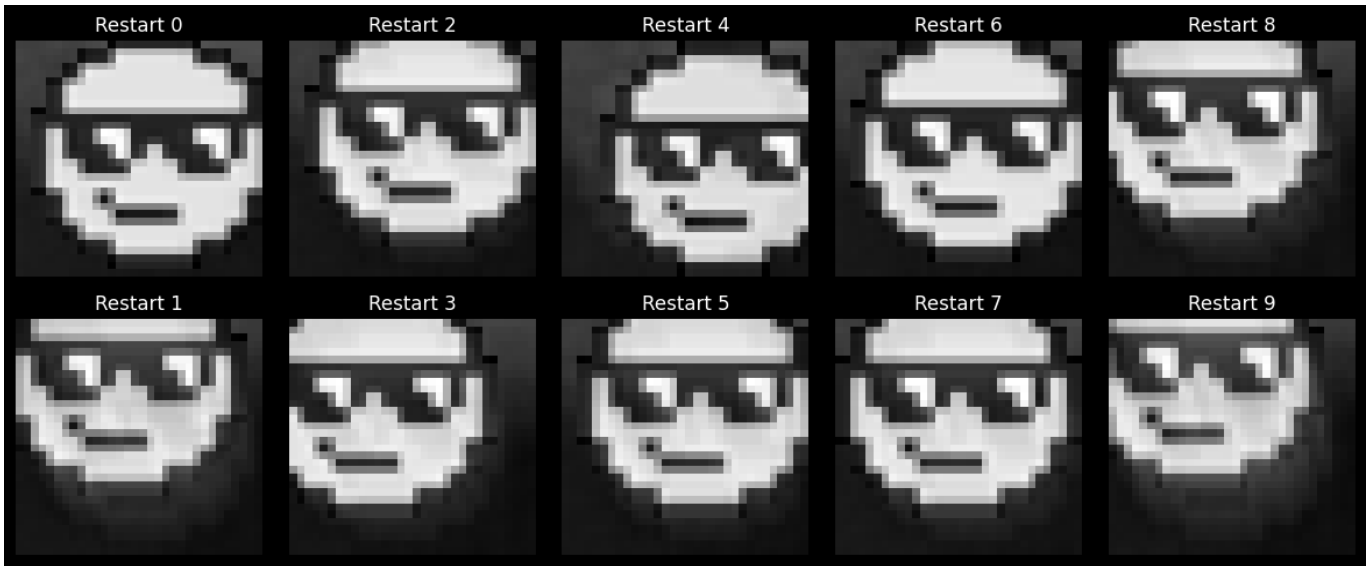

Рис. 2: Задетектированные лица

И вот такие вот картинки фона:

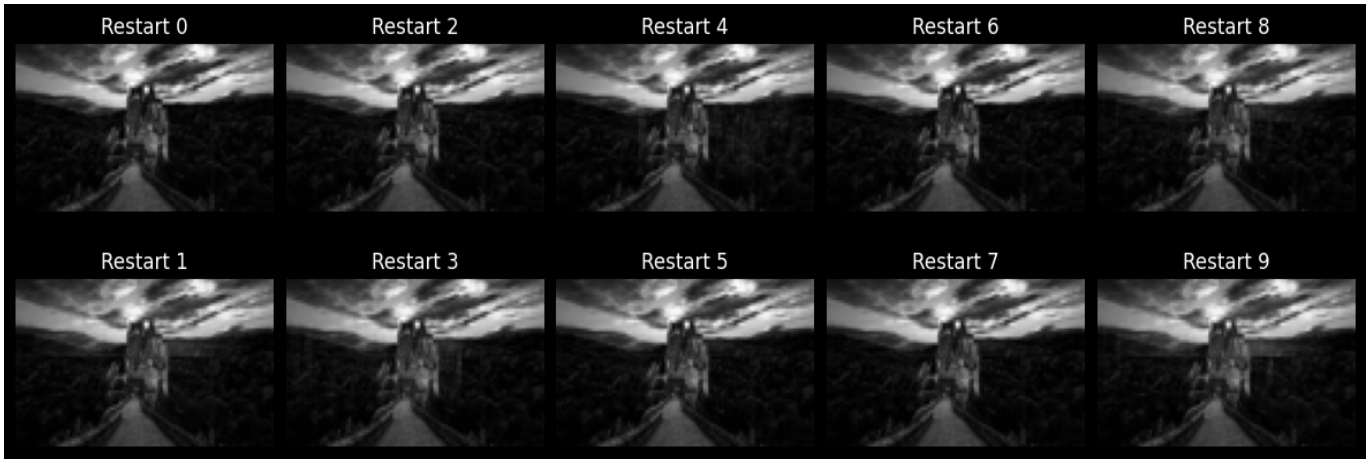

Рис. 3: Найденные фоны

По фону влияние начального приближения параметров на результаты оценить довольно сложно, дифф почти незаметен, а вот по лицу мы видим, что бывает задетектировано не все лицо, оно бывает бледнее или смазаннее, а бывает и ситуация, как в рестарте 6, то есть довольно четкое и практически полное изображение лица. И еще лицо меняет свою позицию. Так что уже можно сказать, что есть смысл запускать EM несколько раз для получения более хорошего результата, вернее, конкретно для этой задачи однозначно стоит запускать EM алгоритм из разных начальных приближений.

И есть еще график нижней оценки логарифма правдоподобия для различных начальных приближений:

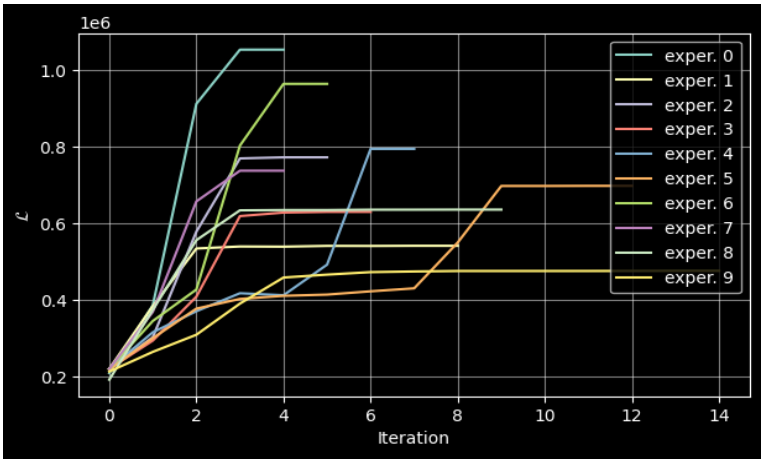

Рис. 4: Нижняя оценка логарифма правдоподобия для разных начальных приближений

И тут тоже довольно легко понять, что раз оптимизация идет итеративно сначала по одной части параметров, потом по другой, можно сойтись к локальному оптимуму и там остаться, а глобальный так и не найти.

3.2 Влияние размера выборки и уровня зашумления на результаты работы

Запустите EM алгоритм на сгенерированных выборках разных размеров и с разным уровнем зашумления. Изучите, как изменения в обучающей выборке влияют на результаты работы (получаемые F, B и $L(q, \theta, A)$). Оцените при каком уровне шума EM-алгоритм перестает выдавать вменяемые результаты. Для сравнения значений $L(q, \theta, A)$ для выборок разного размера стоит нормировать его на объем выборки.

Вот результат запуска:

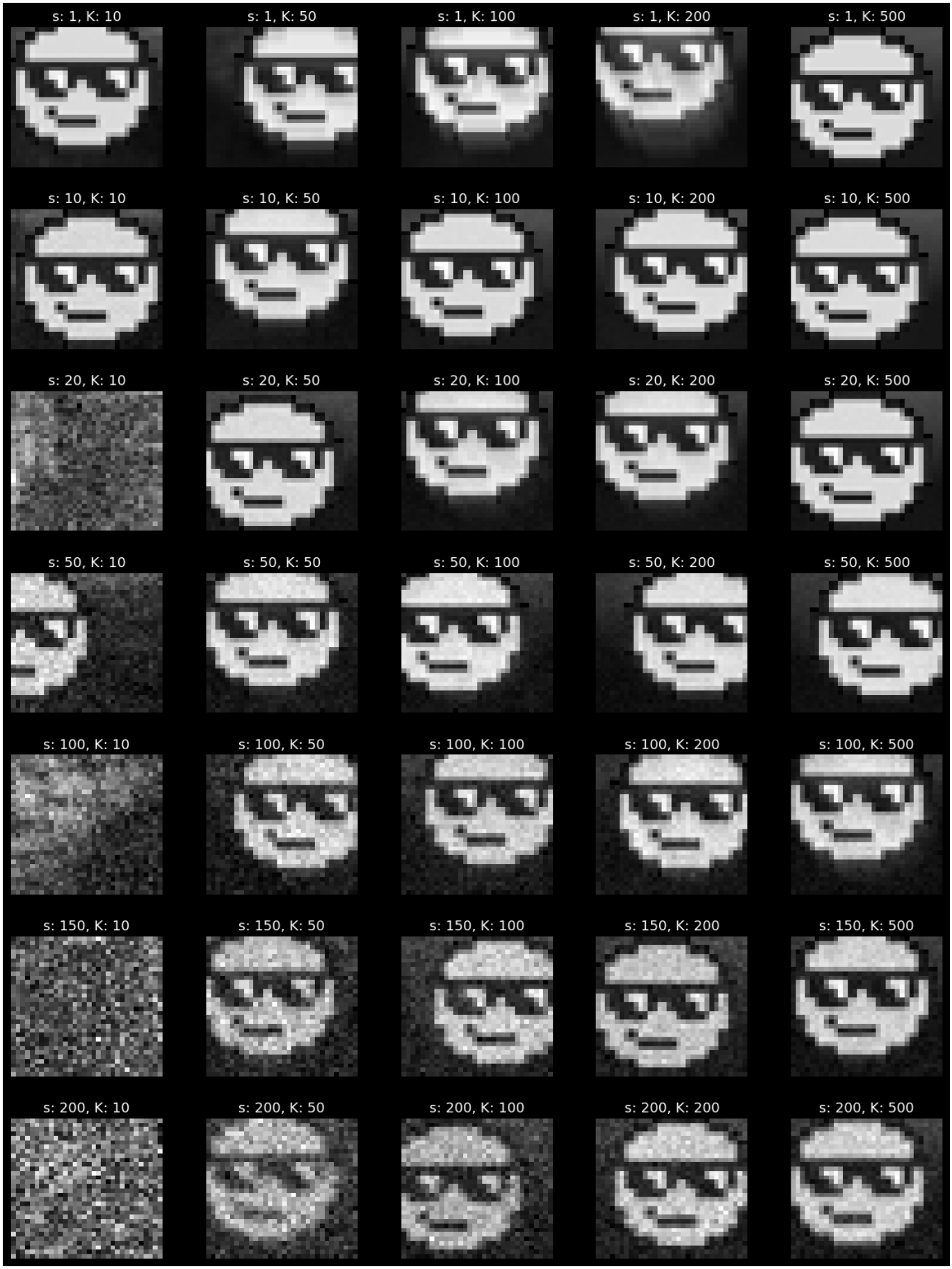

Рис. 5: Лицо с разным уровнем шума и разным размером выборки

Рис. 6: Фон с разным уровнем шума и разным размером выборки

Общее наблюдение по лицу и фону: чем меньше шум и больше выборка (больше картинок), тем лучше будет результат работы алгоритма, иначе - наоборот, много шума при маленькой выборке не дает нам нормально справиться с задачей.

Теперь по графику для $L(q, \theta, A)$:

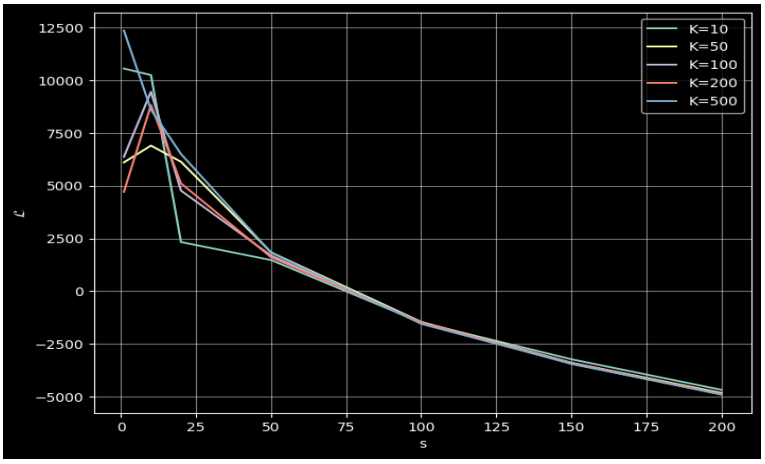

Рис. 7: $L(q, \theta, A)$ для разных s и K

Я отнормировала $L(q, \theta, A)$ на размер выборки, как и просят в задании. Получилось, что значение $L(q, \theta, A)$ падает при увеличении уровня шума.

На моей выборке (я брала размеры K до 500) вышло так, что уже при s = 200 нужны значения K примерно больше 100, чтобы получить относительно четкое изображение. А вообще и при уровне шума 20 алгоритм может не выдать результат на маленькой выборке. Вот такие дела.

3.3 Сравнение качества и времени работы EM и Hard EM

Сравните качество и время работы EM и Hard EM на сгенерированных данных. Объясните, почему разница в результатах работы так заметна.

Если рассуждать исходя из теории, то hard EM должен работать быстрее обычного EM, потому что обычный EM использует для своих рассчетов гораздо больше данных, опять же, из этого следует, что hard EM результат будет выдавать хуже, чем обычный EM, потому что у обычного больше информации. Получается некий трейдофф.

На картинках ниже обычный EM и hard EM. И очень красиво получается, что и правда hard EM работает сиииииильно быстрее. При этом да, хотя мои данные довольно простые, hard EM заметно хуже справляется с задачей, например, при s = 1 и K = 10 обычный EM смайлик нашел очень хорошо, а у hard EM получился непонятный шум и даже скорее кусок фона... Похожая ситуация и при s = 50 и K = 500. Параметры в обе модели всегда подавались одинаковые, за исключением use_MAP.

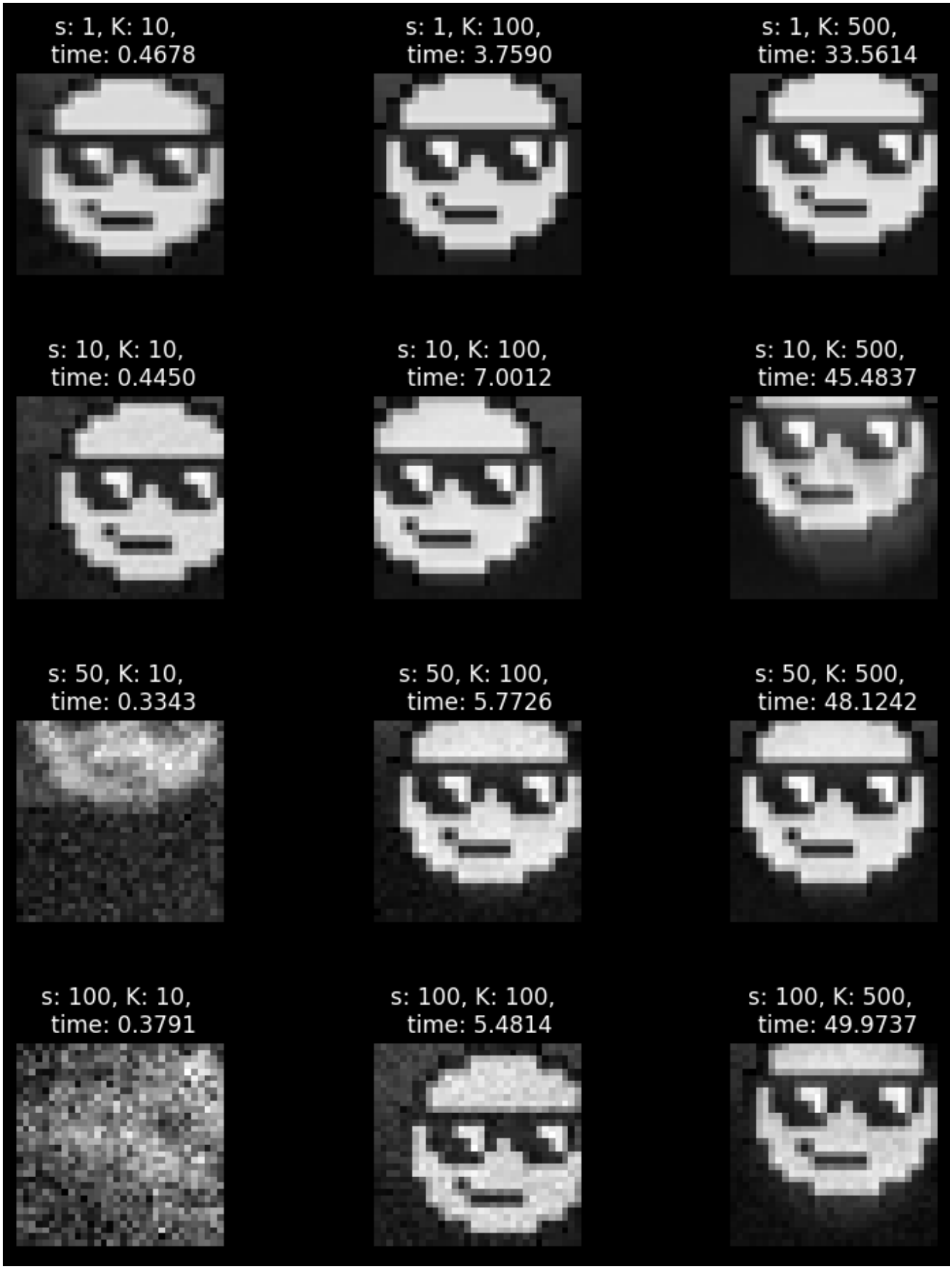

Рис. 8: Обычный EM на сгенерированных данных

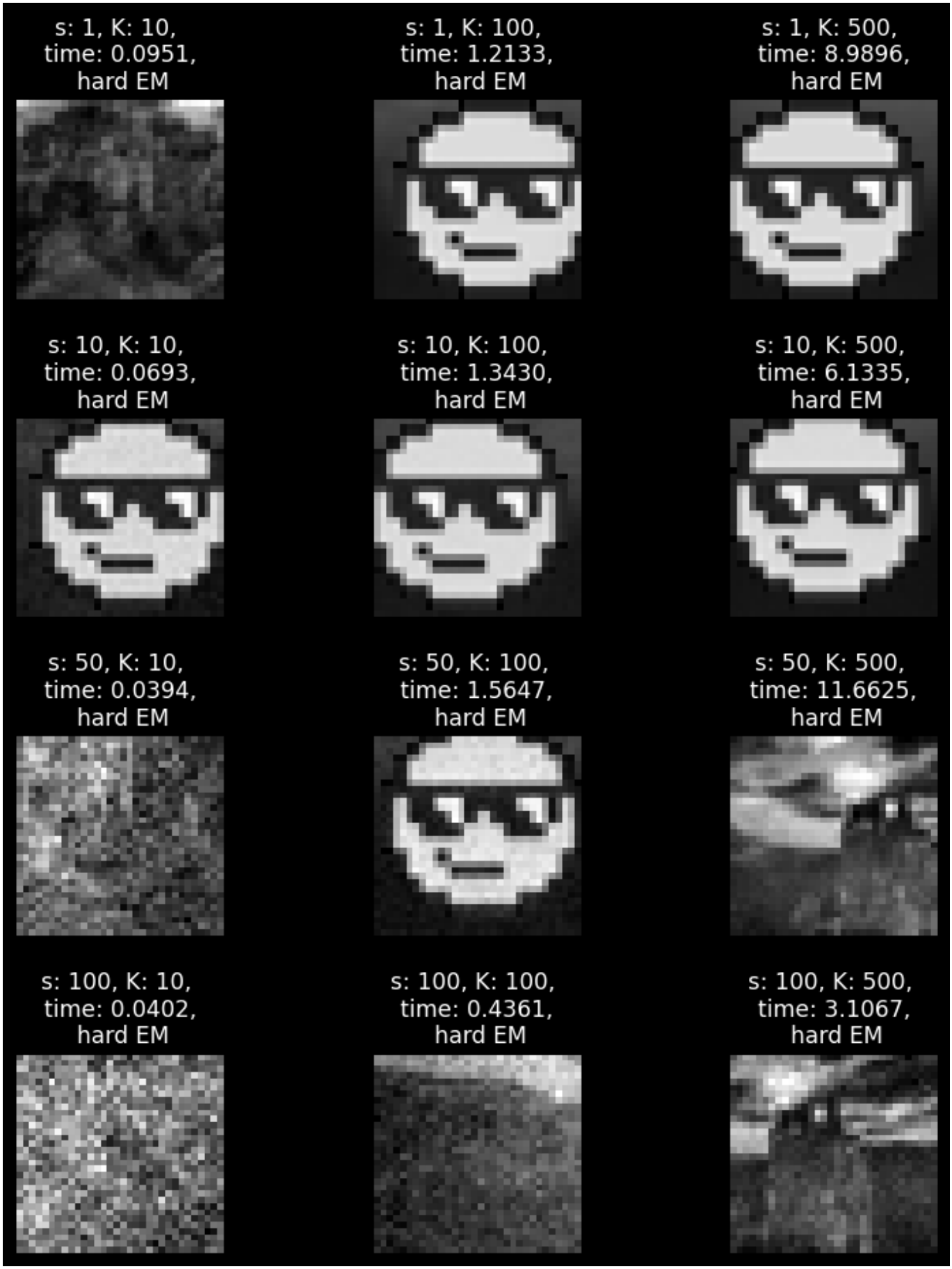

Рис. 9: Hard EM на сгенерированных данных

3.4 Результаты работы на реальных данных

Примените EM алгоритм к данным с зашумленными снимками преступника. Приведите результаты работы алгоритма на выборках разного размера.
Не идеал, но даже так очень неплохо, можно распознать лицо и фон.

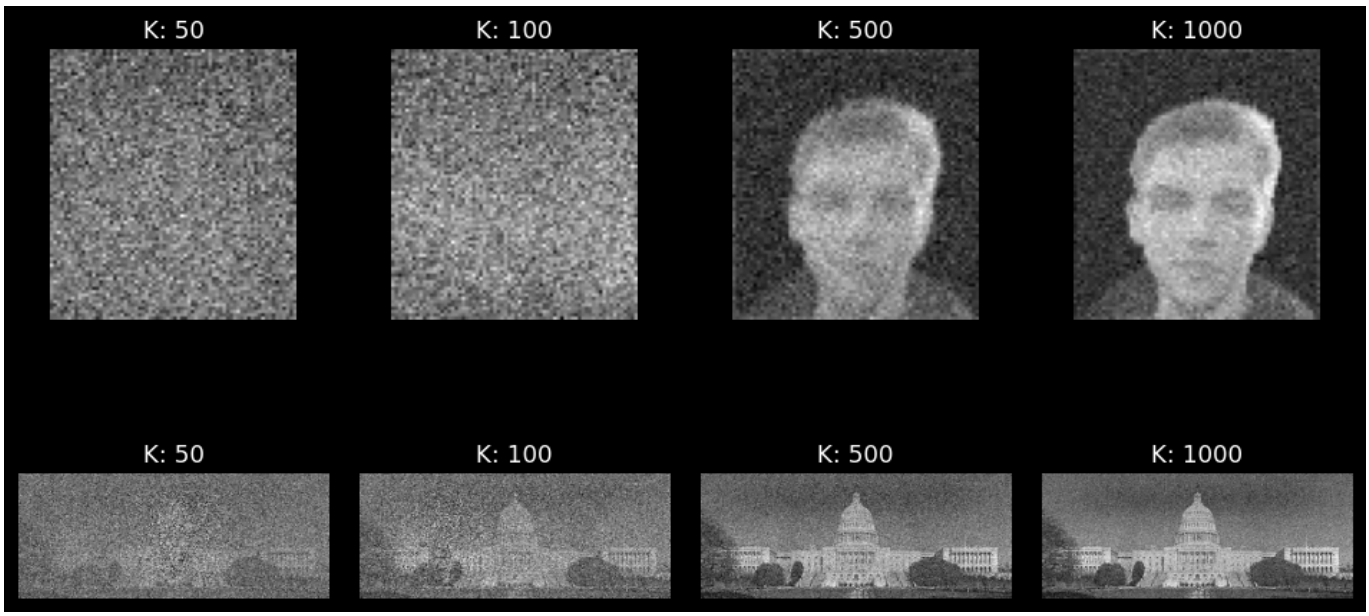

Рис. 10: Результаты работы алгоритма на реальных данных на выборках разного размера

3.5 Модификация EM алгоритма

Предложите какую-нибудь модификацию полученного EM алгоритма, которая бы работала на данной задаче качественнее и/или быстрее.

Итак, идеи:

- Можно умнее ставить начальные приближения параметров, так как они все-таки заметно влияют на успешность процесса. Для начального фона брать среднее изображений по датасету (как будто бы это очень неплохая стратегия), мы так буквально будем изначально аппроксимировать фон. Я даже потестила идею. Вот результат:
- Можно при рестарте инициализировать F и B уже не случайным образом, а каким-нить средним предыдущих полученных F и B.