ForgetMeNot: Memory-Aware Forensic Facial Sketch Matching

Shuxin Ouyang[†,§] Timothy M. Hospedales[§] Yi-Zhe Song[§] Xueming Li[†]

[†]Beijing University of Posts and Telecommunications [§]Queen Mary University of London, UK

{s.ouyang, t.hospedales, yizhe.song}@qmul.ac.uk lixm@bupt.edu.cn

Abstract

We investigate whether it is possible to improve the performance of automated facial forensic sketch matching by learning from examples of facial forgetting over time. Forensic facial sketch recognition is a key capability for law enforcement, but remains an unsolved problem. It is extremely challenging because there are three distinct contributors to the domain gap between forensic sketches and photos: The well-studied sketch-photo modality gap, and the less studied gaps due to (i) the forgetting process of the eye-witness and (ii) their inability to elucidate their memory. In this paper, we address the memory problem head on by introducing a database of 400 forensic sketches created at different time-delays. Based on this database we build a model to reverse the forgetting process. Surprisingly, we show that it is possible to systematically "un-forget" facial details. Moreover, it is possible to apply this model to dramatically improve forensic sketch recognition in practice: we achieve the state of the art results when matching 195 benchmark forensic sketches against corresponding photos and a 10,030 mugshot database.

1. Introduction

Facial sketch recognition is an important law enforcement tool for determining the identity of criminals where only an eyewitness account of the suspect is available. In this situation, a forensic sketch artist renders the face of the suspect by hand or with compositing software based on eyewitness description. The facial sketch is then disseminated in the media, but the crucial capability is to then identify the suspect by matching it against a photo mugshot database.

Motivated by this, the computer vision [12] and biometrics [2] fields have extensively studied sketch to photo face matching. However, practical matching of forensic sketches to photo databases remains an unsolved question. This is because studies have primarily focused on matching viewed sketches rather than the rarer forensic sketches. *Viewed sketches* such as those in the popular CUHK [23] database are drawn by artists while viewing a photo. As such there is no forgetting issue, and the sketches are accurate renditions of the subject. The cross-modal sketch-photo gap is thus small, and viewed sketches are relatively easy to match – resulting in benchmark performance saturated at near-perfect [1, 2, 4, 12]. *Forensic sketches* are drawn based on eyewitness description, possibly days after the event. Despite being the practically relevant variant of the problem for law enforcement, forensic sketch matching remains both relatively unstudied and unsolved. It is a much harder and unsolved problem due to the sketch-photo gap being widened by: (i) forgotten / inaccurate memory of facial details [7], and (ii) imperfect communication of memory [5] (whether to a human sketch-artist or software compositor [7]). Nevertheless, it is relatively unstudied largely due to lesser availability of forensic sketch benchmark databases, which is why we introduce a new forensic sketch database.

In computer vision, facial sketch-photo matching has been studied extensively using a variety of approaches including invariant feature engineering [1, 2, 4, 12], cross-modal regression/synthesis [22, 23] and shared subspace learning [20]. These contributions address the sketch/photo modality gap, but do not address the issue of forgotten or inaccurately remembered details due to imperfect memory. In contrast, psychology [25] and forensic psychology [6] have studied the reliability of different facial features in human face matching, and the fading of memory with time [7]. This has provided some insights into human recognition (internal facial features are more important overall), and the reliability of human memory, for example that memory fidelity drops rapidly after a few hours [7]. This means that forensic sketches are very inaccurate in practice, because they are usually taken days after the event [6, 7]. Thus the memory gap is the key underlying problem to solve.

Motivated by these studies in human memory and recognition, we investigate here whether it is possible to bring learning and computer vision techniques to bear to ameliorate the memory gap problem. To disentangle the three factors (cross-modal, forgetting, and imperfect communication) in the forensic sketch/photo gap, we introduce a new

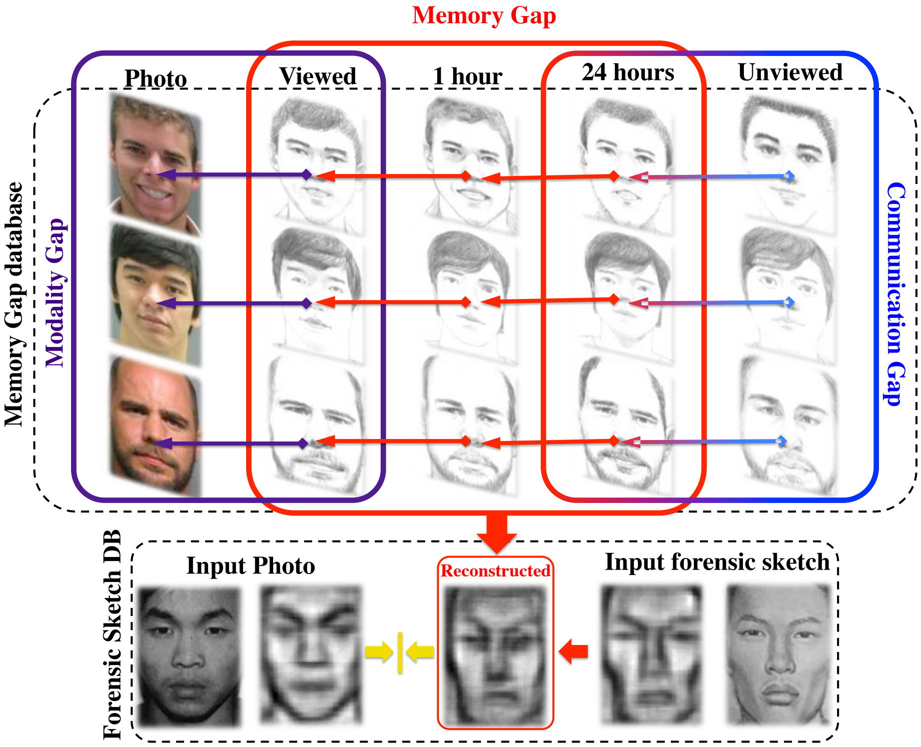


Figure 1. Database and approach overview. We first learn a projection for "un-forgetting", as well as modality and description gap (top). We apply this projection to improve (un-forget) forensic sketches before matching against photos (below). Reconstructed sketch (red) is a closer match to the true photo (bottom left) than the input forensic sketch (bottom right) (visualisation with HOGgles [21]).

facial sketch memory gap database that contains 100 subjects. Uniquely, each subject has a photo, a viewed sketch, a 1-hour delay sketch, a 24-hour delay sketch and an unviewed sketch. Based on this database, we investigate the question of whether memory transience is *random* (i.e., all memory errors are equally likely), or there is any *systematicity* in the forgetting process (i.e., misremembered details occur with some kind of predictable pattern that can be exploited). Somewhat surprisingly, we demonstrate that it is possible for a machine learning model to input a forensic sketch, and to some extent reverse the forgetting process to produce a more accurate sketch that is easier to match.

Based on our memory gap database and model, we aim to improve forensic sketch to mugshot matching: by modelling the photo-sketch modality gap, imperfect communication gap and – uniquely – by modelling a map from memories of old to recently seen faces to correct misremembered facial details. Since forgetting dynamics differ across time periods [7], it is unclear how to model the memory gap data: a single model covering forgetting across different time-periods is too coarse, but a distinct model of the forgetting in time-slice of the database is too specific. Similarly, the overall forensic sketch matching task spans modality, communication and memory gaps. An intuitive approach would therefore be to apply *in sequence* multiple models trained to span each of these gaps. We show that while this is effective, a better solution in practice is to apply multi-task learning [24] to build a single model trained to span the longer 24h memory gap, but with the others (short-term memory, modality and communication) as auxiliary tasks. Finally, we demonstrate the practical value of these contributions by transferring the model learned on our memory gap database to a realistic forensic task [11, 12] of matching 195 forensic sketches against corresponding photos and a 10,030-mugshot database. The results demonstrate a large improvement over the previous state of the art. An overview of our proposed framework is illustrated in Figure 1.

2. Related work

Facial sketch-photo recognition: Studies on matching facial sketches to photos can be classified based on the type of sketches used: *viewed*, *semi-forensic* and *forensic*, and whether the sketches are hand drawn, or computer composited. The majority of previous studies have focused on viewed sketches due to being an easier task with accessible benchmark databases. Representative approaches to viewed sketch recognition include bridging the gap with MRF-based photo-sketch synthesis, [23], learning common subspace for comparison with PLS [20], or engineered new invariant descriptors [8]. For further details, we refer the reader to the survey in [17]. Recognition rates on the main viewed sketch benchmarks [23] have reached 100% [8], so viewed sketch recognition can be considered solved.

Forensic sketch face recognition: One of the earliest studies to discuss automatically matching forensic sketches with photos was [10]. It highlighted the importance, as well as complexity and difficulty of forensic sketch based face recognition. The first significant demonstration of automated forensic sketch matching was [12], which combined feature engineering (SIFT and LBP) with a discriminative (LFDA) method to learn a weighting that maximised identification accuracy. Later studies such as [2] improved these results, again combining feature engineering (Weber and Wavelet descriptors) plus the discriminative learning (genetic algorithms) strategy to maximise matching accuracy.

Unlike viewed sketches, forensic sketch databases are few and small in size. The main sketch/photo databases are 159 pairs identified by [12], and 190 pairs in the IIIT-D database [2]. A realistic evaluation of sketch-based face matching should also include a large pool of mugshots to match against, in addition to the true photo corresponding to each sketch. Despite this, only a few studies have evaluated forensic sketch matching algorithms in this way. Notably [12], which trained a matching model on viewed sketches and then tested matching 159 forensic sketches against corresponding photos and a 10,030 mugshot database. In this paper we also evaluate our approach in this rigorous way, and show that the results can be significantly improved by explicitly modelling the human visual memory components.

Regression models: Regression models are widely used in cross-domain face recognition [17]. For facial sketch matching, regression models may provide facial sketch↔photo synthesis [22] to support matching, for example via support vector regression (SVR) [26]. Alternatively, Partial Least Squares (PLS) models may be used to map images in each modality to a common subspace where they are more comparable [20]. Although widely and effectively used, all prior work has focused on regression modelling to tackle the modality-gap problem rather than the memory-gap problem. In this paper, we exploit Gaussian Process regression to deal with both the memory-gap and the modality-gap components in forensic sketch matching.

Facial Attributes: Study of facial attributes [14, 16] is a topical problem in computer vision. It is also relevant to forensic sketch recognition because encoding sketches and photos in terms of facial attributes can help to bridge the sketch/photo modality gap [18], or prune the matching space [12]. However, attributes are vulnerable to forgetting as well, so the attributes of a sketch may mismatch those of the corresponding photo even if they are perfectly detectable by computer vision techniques.

Human memory and forensic sketches: Studies have shown the ability of individuals to recognise faces depends on different facial features according to the level of familiarity [25]. Internal facial features are important for identification of familiar faces, and external features for unfamiliar faces [6]. It remains to be seen if/how these findings translate to automatic face recognition, so we use whole face images in our study. With regards to the forgetting process, forensic psychology studies have found that memory fidelity drops dramatically between the first hour and first 24 hours after witnessing a face. However, in practice forensic sketches are rarely made within the first day [7]. Thus, any mechanism capable of bridging this gap automatically is expected to both have a large impact on quantitative recognition performance and forensic police work in practice.

Contributions: Overall, our contributions are as follows: (i) We present a new memory gap facial sketch database with 100 subjects each with a photo and four sketches that disentangle different aspects of the forensic sketch gap (400 sketches in total). (ii) We use this database to demonstrate that there is systematicity in facial forgetting, by showing that inaccurate forensic facial sketches can be automatically improved by machine learning methods trained to recover 'recent' from 'old' face memories. (iii) We transfer the learned memory reconstruction models to a realistic forensic sketch matching benchmark. The results significantly outperform the previous state of the art [11, 12, 15] at matching forensic sketches against corresponding photos and a large 10,030 mugshot database.

3. Memory-Aware Facial Sketch Modeling

The forensic sketch-photo matching task is complicated by three distinct challenges. Photo/sketch modality change, forgetting, and communication (of memory to sketch artist/compositing software) issues all contribute. We create a dataset designed to disentangle these issues. It contains N subjects, with photos $D^p = \{\mathbf{x}_i^p\}_{i=1}^N$ and sketches drawn with different conditions $D^s = \{\mathbf{x}_i^t\}_{i=1}^N$, t = (v)iewed, (1) hour, (24) hour and (u)nviewed. Each image is assumed to be represented by a d-dimensional feature vector $\mathbf{x}$. The task of nearest-neighbour (NN) matching a viewed sketch $\mathbf{x}^{t=v}$ to a photo database would be

$$i^*_{NN} \quad = \quad \operatorname*{argmin}_i |\mathbf{x}^v - \mathbf{x}_i^p| \, . \qquad (1)$$

Studies focusing on bridging the modality gap by linear regression-based synthesis or linear subspace projection aim to solve a similar task, after learning a suitable regression matrix W^v or projections W^v and W^p respectively:

$$i^*_{map} \quad = \quad \operatorname*{argmin}_i |W^v \mathbf{x}^v - W^p \mathbf{x}_i^p| \, . \qquad (2)$$

Memory Modelling: Making use of our memory-gap database, we can separate contributing components of the forensic-sketch gap. For example, training $W^{v \to p}$ in

$$W^{v \to p} = \underset{W^{v \to p}}{\operatorname{argmin}} \sum_i \|\mathbf{x}_i^p - W^{v \to p}\mathbf{x}_i^v\|_2^2 \quad (3)$$

is the conventional task of learning to bridge the modality gap between photos and viewed sketches. Training $W^{u \to v}$ would be learning to correct the communication gap. While training $W^{24 \to v}$ in

$$W^{24 \to v} = \underset{W^{24 \to v}}{\operatorname{argmin}} \sum_i \left\|\mathbf{x}_i^v - W^{24 \to v}\mathbf{x}_i^{24}\right\|_2^2 \quad (4)$$

is learning to *correct 24 hours worth of transience*, independent of the modality or communication gap. Given the conditions in our memory-gap database, there are a variety of potential tasks (10 in total) including: correcting the modality $v \to p$ or short term memory gap $1 \to v$; reducing or completely correcting the long-term memory gap $24 \to 1$ or $24 \to v$ respectively; and full forensic sketch matching $u \to p$ (see Sec. 5.1 for full list). We will learn all 10 tasks allowed by our database.

Mapping Strategy: Rather than the most common linear projection approach to these learning tasks [20], we use Gaussian Process Regression (GPR) [19]. We take this approach because: (i) GPR provides a more flexible non-linear mapping, and importantly (ii) as a Bayesian regression framework, GPR provides a distribution over the reconstruction rather than a single point estimate. This uncertainty metric at each point of the reconstruction turns out to be important to improve matching performance, by automatically weighting each feature according to its reliability.

Exploiting Multiple Models: As mentioned earlier, our memory-gap database provides 10 potential modelling tasks. The most obvious ways to use these for practical forensic sketch matching would be: (i) apply the model learned for direct forensic sketch-photo matching $u \to p$, or (ii) given multiple models trained to correct the different sources of error, sequentially apply them to correct each source of error in turn, e.g., $u \to 24 \to 1 \to v \to p$.

Clearly some of these tasks are related (e.g., tasks $1 \to v$, $24 \to 1$, $24 \to v$ span different steps of forgetting). So an alternative approach that will turn out to be better is to learn all the tasks together in a multi-task learning framework. In this way each task shares information with – is regularised by – the others. Specifically, we will jointly learn the tasks with Multi-Task Gaussian Process Regression (MTL-GPR).

3.1. Improving Forgotten Faces with MTL-GPR

Single Task Modelling: GP regression can be applied to cross-modal/memory-gap problems such as those in Eqs. 2-4, but learning a non-linear projection. Denoting now features in input and target conditions as x and y respectively, our database provides training pairs $D = \{\mathbf{y}, \mathbf{x}\}$. For any query point x^* the GPR prediction for y^* is:

$$p(y^*|x^*, D) \sim \mathcal{N}(\mathbf{k}_*^T K^{-1}\mathbf{y}, \mathbf{k}_{**} - \mathbf{k}_*^T K^{-1}\mathbf{k}_*) \quad (5)$$

where matrix K is the covariances at all pairs of train points, vector $\mathbf{k}_*$ is the train-test covariances, $\mathbf{k}_* = [\kappa(x_*, x_1)...\kappa(x_*, x_N)]$ and $k_{**} = \kappa(x_*, x_*)$. We take the most common squared-exponential kernel $\kappa(x, x') = \exp(-\frac{1}{2l^2}(x - x')^2)$, and the kernel hyper parameter l can be tuned by gradient on the marginal likelihood [19].

Multi Task Modelling: In our problem there are 10 distinct mapping tasks, which we learn together in a MTL-GPR framework. Following [3], we learn GP regression with predictions for tasks l and k correlated as:

$$< f_l(x) f_k(x') > = K_{lk}^f \kappa(x, x') \quad (6)$$

Here l and k index any two conditions in our memory-gap database, and K^f is the 10×10 PSD matrix of inter-task similarities. Standard GP predictions can then be made using this covariance. Importantly, with this approach, the key task similarity matrix K^f can also be learned along with the kernel hyper parameters l via the marginal likelihood [3].

3.2. Matching Forgotten Sketches to Photos

Correcting Inaccurate Memory: For any task provided by our database, reconstruction is performed by computing the GP posterior of each target feature. For example, to improve an unviewed sketch $u \to v$, we would compute the predictive distribution $p(\mathbf{x}_*^v|\mathbf{x}_*^u, D) \sim \mathcal{N}(\mu_{\mathbf{x}_*}, \sigma_{\mathbf{x}_*}^2)$, as given by Eq. 5. The new sketch would then be given by the mean of the posterior normal $\mu_{\mathbf{x}_*}$, and the confidence of each feature dimension by the corresponding variance $\sigma_{\mathbf{x}_*}^2$.

Matching across Memory or Domain Gap: With this framework matching can be performed by calculating the likelihood of each mugshot in the gallery under the posterior predictive distribution of the probe sketch. For example, after training on our memory gap database D, we can use model $u \to p$ to match a forensic sketch $\mathbf{x}_*^u$ against a database of mugshots $X^p = \{\mathbf{x}_i^p\}_{i=1}^N$ as follows:

- Compute the distribution over the expected photo corresponding to the forensic sketch: $p(\mathbf{x}^p|\mathbf{x}_*^u, D)$.
- Pick the photo with maximum likelihood under this predictive photo distribution: $i^* = \underset{i}{\operatorname{argmax}} p(\mathbf{x}_i^p|\mathbf{x}_*^u, D)$.
- In practice, we model each dimension of the target independently with GPR, so this is equivalent to $i^* = \underset{i}{\operatorname{argmax}} \sum_k (x_{ik}^p - \mu_{x_{*k}})^2/\sigma_{x_{*k}}^2$. Where x_{ik}^p, $\mu_{x_{*k}}$ and $\sigma_{x_{*k}}^2$ respectively are the $k-th$ dimension of the target photo, posterior predicted photo mean and variance.

4. Memory gap database

In this section we describe our memory gap database and its creation procedure in more detail[1]. 100 subjects are chosen from mugshots.com, which releases mugshots of real criminals. For each subject one frontal face photo is selected, and four types of sketches are drawn:

Viewed: Sketches are drawn while the artist looks directly at the mugshot photos.

1 hour: Mugshot photos are viewed by the artist, and sketches are drawn one hour later. Thus, compared to viewed sketches, the sketch is 'corrupted' by one hour worth of memory transience.

24 hours: Mugshot photos are viewed by the artist, and drawn 24-hours later.

Unviewed: Sketches are drawn by an artist based on the description of an eyewitness who has seen the mugshot photo immediately before (but does not view it during the sketching). The artist does not see the photo. In this case, the *memory gap* is negligible, but it is the only condition in the database where the *communication gap* of imperfect communication between the eyewitness and artist exists.

The reason for this design of the collection procedure is so that the modality and communication gaps can be isolated (in photo-viewed and viewed-unviewed respectively) from the memory gap (24h to 1h to viewed). This potentially enables specific models to be built to address each contributing factor of the forensic sketch challenge.

To build the memory gap database, over 20 art students are selected to contribute as both sketch artists and eyewitness. Each artist is asked to draw all four kinds of sketches for each subject. This way the sketches for each mugshot do not have inter-artist variability, but the drawing order is such that forensic sketches are fully unviewed.

5. Experiments

5.1. Datasets and Settings

Databases: We study three databases: The contributed **Memory Gap Database (MGDB)**, where we have also annotated each image with 40 binary facial attributes from the ontology provided by [18]; a **Forensic Composite Database** with 51 forensic composite-photo pairs [7], and the **Forensic Sketch and Mugshot Database (FSMD)**. The latter consists of two parts: 195 forensic sketch-photo pairs [2, 12] and a large background gallery of mugshots to search against, in order to replicate a real-world scenario where a law-enforcement agency would query a large gallery of mugshot images with a forensic sketch. We use the same 195 sketch-photo pairs as [12, 18]. The mugshot gallery used by [11, 12] was not released publicly, so we simulate this as best as possible by downloading 10,030 mugshots from mugshots.com (the same source used by [12]).

Memory-Aware Model Training: All sketch and photo conditions (t=photo, viewed, 1 hour, 24 hour and unviewed) are used to exhaustively construct the 10 possible reconstruction tasks. For each task, sketches corresponding to two-thirds of subjects serve as training data, and the others serve as testing data. The 2/3s training subjects and 10 tasks are used to jointly train 10 models via MTL-GPR. We explore performance on the testing split of Memory Gap Database, before transferring to FSMD for final evaluation.

Overall ten regression tasks were trained: 1) viewed sketch to photo, 2) 1 hour sketch to photo, 3) 24 hour sketch to photo, 4) unviewed sketch to photo, 5) 1 hour to viewed sketch, 6) 24 hour to viewed sketch, 7) unviewed to viewed sketch, 8) 24 hour to 1 hour sketch, 9) unviewed to 1 hour sketch and 10) unviewed to 24 hour sketch. Some of these are illustrated in Fig. 1.

Features and settings: We normalise all photo and sketch images to 256×196 and align them by normalising on inter-ocular distance. Each image is then represented with HoG features. We compute dense HoG feature over a regular grid (16×16 step size), which results in a feature vector of dimension 5,952 for each image. For each image, 40 attributes are also detected using SVM detectors trained using the ground-truth attributes on the training split [18].

Baselines: In addition to our MTL-GPR memory-aware model, we also consider alternative regression methods that could potentially model the gaps across database contexts:

Nearest Neighbour (NN): Direct matching. Ignore the gap.

Linear Regression (LR): Linear (L2 regularised) regression is the simplest explicit mapping approach.

Polynomial Support Vector Regression (SVR): SVR was used in [26] to accomplish sketch-photo synthesis.

Polynomial Multi-Task Learning: We use the [24] implementation of the popular GO-MTL [13] multi-task learner. By exploiting task relatedness, this may perform better than SVR. In initial experiments we found polynomial MTL significantly better than linear, so we report the former.

(Single Task) Gaussian Process Regression (GPR) [19]: Compared to the others, GPR provides a non-parametric probabilistic prediction with an estimate of uncertainty that can be used for matching as in Sec 3.2.

Sequential GPR: As mentioned in Sec 3, this is the intuitive baseline of applying a number of the 10 GPR models in sequence to correct distinct error sources.

5.2. Memory-Aware Model Analysis

In this section, we analyse the MTL-GPR reconstruction of faces, as represented by HoG features[2]. To help inter-

[1] Available to download at http://sketchx.eecs.qmul.ac.uk/downloads.html

[2] The analysis could in principle be done with pixels, but this would be computationally expensive due to higher dimensionality.

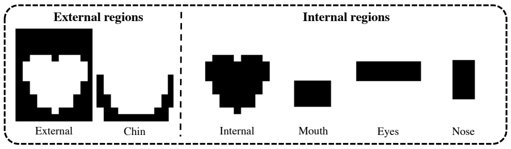


Figure 2. Illustration of facial regions.

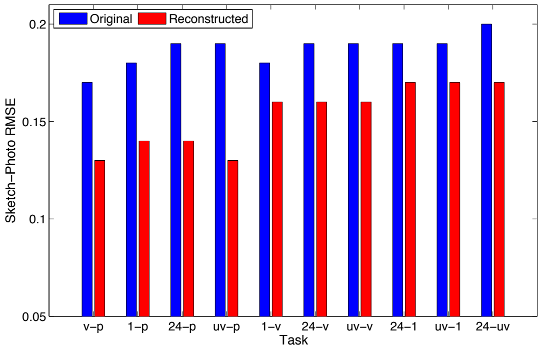


Figure 3. Learned reconstruction reduces sketch/photo gap for each task in MGDB database: RMSE averaged across full face.

Table 1. RMSE of sketch/reconstruction vs photo according to regions, averaged across all ten tasks in MGDB.

Region	Photo v.s. Original Sketch	Photo v.s. Projected Sketch
External	0.20 ± 0.013	**0.16**± 0.025
Chin	0.20 ± 0.014	**0.16**± 0.023
Internal	0.18 ± 0.003	**0.16**± 0.015
Mouth	0.17 ± 0.007	**0.16**± 0.012
Eyes	0.18 ± 0.003	**0.15**± 0.023
Nose	0.18 ± 0.011	**0.14**± 0.018

pret the results, we also divide the facial hog feature maps into external regions and internal regions: external, internal, eyes, nose, mouth and chin [25], as shown in Fig. 2. To investigate whether our memory model helps to bridge the gap between photo and forensic sketch, we calculated RMSE between sketch/reconstructed sketch and the corresponding photos. The results are shown broken down by facial region and averaged over tasks (Tab. 1) and averaged over all regions broken down by tasks (Fig. 3). From these we can see that: (i) Each learned projection task in the MGDB database reduces the sketch-photo RMSE. (ii) This demonstrates that sketches drawn at different delays contain some systematic shift that it is possible to reverse, or it would not be possible to learn a model that consistently improves RMSE. (iii) Reconstruction consistently improves RMSE for each distinct semantic facial region.

5.3. Face matching: Memory gap database.

In this section we quantitatively evaluate face matching performance on the test split of the memory gap database. As outlined in Sec 5.1, we compare a variety of baselines to our proposed MTL-GPR and report the rank 1 (perfect match) accuracy for each of the 10 tasks in Tab. 2. The row and column give the MGDB image pair (training task). The column gives the MGDB sketch input for testing, and the task is always to match against photos using the corresponding training model.

Efficacy of memory-aware models: From Tab. 2, we can draw the conclusions: (i) Sketch reconstruction with linear regression does not consistently improve on direct NN matching, suggesting that a linear projection is insufficient. (ii) Every non-linear approach to bridging the modality/memory gap performs better than direct NN matching with no memory gap model, but among the baseline memory gap models, there is no clear winner or loser. (iii) Our MTL-GPR is the clear winner overall, often with significant margins over the next best (e.g., 87% vs 57% in $24 \to v$ setting). (iv) That MTL-GPR outperforms regular GPR demonstrates that there is common information in each of the distinct tasks that can be extracted and shared. (v) In some cases the gain from an explicit un-forgetting model is vast: In the $24 \to v$ setting, performance triples from 29% to 87% comparing NN matching with MTL-GPR.

Significance of Bayesian Memory Gap Model: One of the reasons for the GP methods' good performance is their ability to account for reconstructed feature reliability in matching (Sec 3.2). We demonstrate this in Tab. 3, where we compare performance with and without the use of the reconstruction variance. Clearly accounting for reconstruction reliability significantly benefits performance.

Qualitative Analysis: The average variance map across the database is shown in Fig. 5(right). The model confidently predicts both internal (eyes, mouth) and external (hair, chin) facial regions [25], while giving less weight to skin regions (forehead, cheeks), where texture may not be predictable from the sketch.

The MTL-GPR framework also aims to discover task relatedness. The learned task relatedness matrix K^f is shown in Fig. 5(left). The clear block structure here shows that the tasks with sketches as target context are much more related to each other than those with photos as the targets. The $24 \to 1$ task is also noticeable as sharing structure with many of the other sketch predictors (cross structure within the block).

5.4. Applying Memory-Aware Models to Forensic Sketch Matching

Matching on Forensic Sketch Database: All ten learned memory-aware models are transferred to the forensic sketch database, which includes 195 forensic sketch-photo pairs. Few experiments have been done on forensic sketch database, except [18] which focused on using attributes to bridge the sketch/photo gap. To compare directly with [18], we evaluate our models on the same 1/3 test split.

The results are shown in Tab. 4, from which we make the following observations: (i) All our reconstruction models perform significantly better than 9% with HoG matching alone, and almost all outperform the 21% of [18]. (ii) Com-

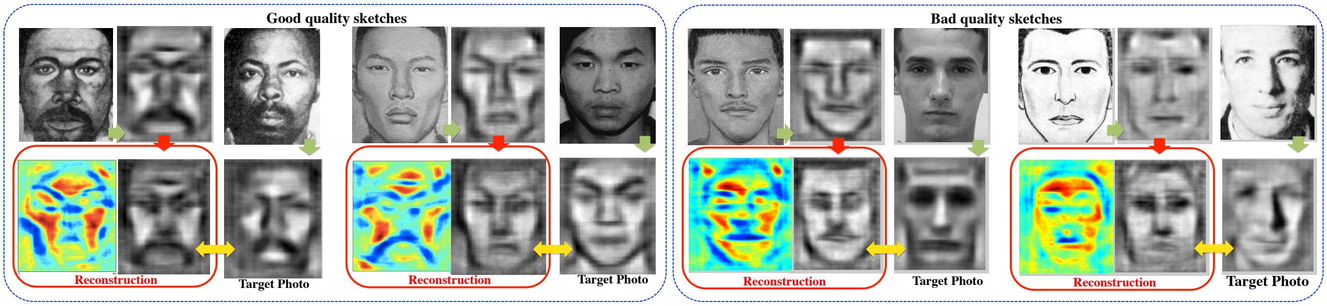


Figure 4. Qualitative results of matching in forensic sketch database. The memory reconstruction model trained on 24 → 1 hour sketches of MGDB is transferred to forensic sketch database. Reconstruction variance improves matching by focusing on reliable features. These good sketches were both retrieved at Rank 1 of 10,225 (10,030+195). Bad sketches were retrieved at Rank 1592 and 1800 respectively.

Table 2. Photo-sketch matching on the memory gap database (Rank 1 accuracy, %). Comparing MTL-GPR, GPR, Polynomial MTL, Polynomial SVR, Linear Regr. and NN. Sketch input is given by column and matched with the model trained on the corresponding cell of MGDB. Average accuracies over 15 random splits of 68 training and 32 testing subjects. See supplementary for standard deviations.

Accuracy	Viewed						1 Hour						24 Hour						Unviewed					
	MG	G-	PM	PS	LR	NN	MG-	G-	PM	PS	LR	NN	MG	G-	PM	PS	LR	NN	MG	G-	PM	PS	LR	NN
Photo	**99**	88	88	90	53	71	**96**	70	65	56	39	51	**90**	55	50	52	32	31	**86**	35	35	38	34	21
Viewed	-	-	-	-	-	-	**90**	58	63	66	52	51	**86**	57	44	46	26	31	**73**	33	32	38	24	21
1 Hour	-	-	-	-	-	-	-	-	-	-	-	-	**69**	41	44	45	26	31	**63**	32	29	35	18	21
24 Hour	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**42**	30	30	32	18	21

Table 3. The importance of Bayesian memory modelling: Rank 1 MGDB match results (%) without/with reconstruction confidence. Average accuracies over 15 random splits of 68 training and 32 testing subjects. See supplementary for standard deviations.

Accuracy	Viewed	1h	24h	Unviewed
photo	86 / 99	85 / 96	60 / 90	50 / 86
Viewed	-	56 / 90	43 / 86	40 / 73
1h	-	-	38 / 69	36 / 63
24h	-	-	-	28 / 42

Table 4. Matching results (Rank 1 accuracy, %) on forensic sketch database (1/3 test split) using MTL-GPR / STL-GPR. Compare: 21% from [18] and 9% by direct HoG matching. Average accuracies over 15 random splits of 68 training and 32 testing subjects. See supplementary for standard deviations.

Accuracy	Viewed	1h	24h	Unviewed
Photo	22 / 35	22 / 34	15 / 40	18 / 41
Viewed	-	65 / 48	40 / 50	33 / 48
1h	-	-	78 / 48	54 / 40
24h	-	-	-	65 / 42

Table 5. Matching results (Rank 1 accuracy, %) on forensic sketch database (1/3 test split) using sequence of STL-GPR models.

$u \to 24$	$u \to 24 \to 1$	$u \to 24 \to 1 \to v$	$u \to 24 \to 1 \to v \to p$
54	28	20	13
$24 \to 1$	$24 \to 1 \to v$	$24 \to 1 \to v \to p$	$1 \to v \to p$
56	39	16	16

paring STL-GPR and MTL-GPR, the models trained with photo targets perform worse when learned jointly, i.e., they suffer negative transfer from the sketch targets. However, the models trained with sketch targets generally perform better, i.e., they successfully share information about bridging the memory gap. (iii) The best model overall is MTL-GPR's 24 $\rightarrow$ 1, suggesting that the biggest single contributor to the forensic sketch gap in practice is the longer term forgetting between 1 and 24 hours. The second best is also memory related 1 $\rightarrow v$.

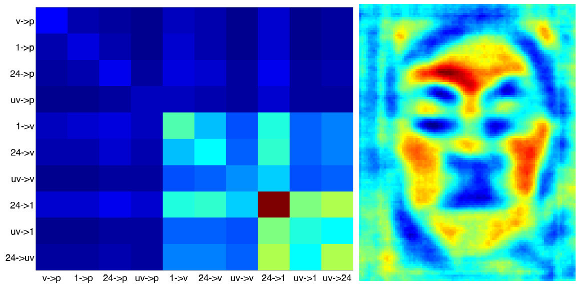


Figure 5. Qualitative results of MTL-GPR model. Left: Estimated task relatedness K^f. Right: Average reconstruction variance.

An intuitive alternative way to exploit the tasks learned in MGDB for forensic sketch matching is to apply the models *in sequence* to correct the various sources of error in forensic sketches. We conduct this experiment for a variety of possible STL-GPR model sequences (Sec 3). The results in Tab. 5 show that while all outperform the 9% of direct matching, none of the multi-step configurations outperform the best single task of 24 $\rightarrow$ 1. Which is itself outperformed by our MTL-GPR 24 $\rightarrow$ 1 in Tab. 4. Based on this analysis, we focus on the contribution of the two MTL-GPR memory models 1 $\rightarrow v$ and 24 $\rightarrow$ 1, which we denote Early and Late, in the final large-scale benchmark experiments.

Matching on Forensic Sketch and Mugshot Database: We now address the full problem of matching forensic sketches to a large database of mugshot photos. We compare the results of our Early and Late-Memory MTL-GPR models to the results of the state of the art LFDA [12] (who also reported the results of a state of the art commercial sys-

Table 6. State of the art comparison. Accuracy (%) of matching 49 good forensic sketches against corresponding photos and 10,030 FSMD database mugshots. * Not directly comparable, used a different 53 sketch probe set.

Accuracy	Rank 1	Rank 10	Rank 50
MTL-GPR Early-Mem	23	23	33
MTL-GPR Early-Mem+Attr	25	25	35
MTL-GPR Late-Mem	33	33	39
MTL-GPR Late-Mem+Attr	**38**	**42**	**45**
LFDA [12]	17	23	33
LFDA [12]+ gender +race	19	27	45
FaceVACS (reported by[12])	2	4	8
KPS [11]*	4	9	21
Deep Features [9]	2	6	15
DFD [15]	6	13	19

Table 7. Accuracy (%) of matching 51 forensic composites against corresponding photos and 10,030 FSMD database mugshots.

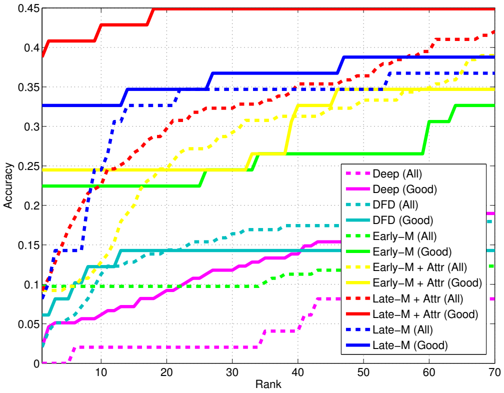
Accuracy	Rank 1	Rank 10	Rank 50
HOG	6	14	20
DFD [15]	2	4	4
MTL-GPR Late-Mem	14	18	26

tem FaceVACS), KPS [11], and DFD [15]. To provide an additional baseline, we also take the best publicly available (photo) Deep face recognition model [9] and use it to extract features for matching. As [12] demonstrated the value of filtering by soft biometrics, we also further combine our models with predicted attributes (trained on memory gap database) with score-level fusion.

In order to compare directly with [12], who break down results by "good" and "bad" quality sketches, we show results in Tab. 6 focusing on a good quality subset of sketches. In Fig. 6, we provide a cumulative match characteristic (CMC) curve, including results for both all 195 sketches as well as the 49 good quality sketches. From the results we can see that: (i) Our memory-gap model significantly surpasses state of the art performance, demonstrating that *the model learned on our database can dramatically improve real forensic sketch matching*, (ii) Of the memory-aware models, the Late-Memory model trained on the 1-24 hour memory gap performs better, reflecting forensic psychology conclusions that the first day's forgetting is significant [7], (iii) Including predicted facial attributes improves performance further, (iv) Using modern deep features with direct matching now outperforms the commercial FaceVACS result, but it is significantly worse than both LFDA [12] and ours: indicating that deep features alone are insufficient to address forensic sketch matching.

Qualitative Examples: Some qualitative examples of our matching process using the forensic database are shown in Fig. 4. Photos and sketches are represented with HoG features (visualised by HOGgles [21]). The learned memory reconstruction model predicts the mean and variance of photo-HOGs. Photos are chosen by their likelihood under the predicted Gaussian distribution, allowing matching to take into account the prediction reliability of each feature.

Figure 6. CMC curves for matching Good (49) / All (195) forensic sketches against corresponding photos and 10,030 FSMD database mugshots.



Matching on Forensic Composite Database: Although our model is trained on sketch rather than software composite faces, we also evaluate whether the learned model is general enough to improve forensic composite matching. Tab. 7 shows the results of retrieving 51 composites from among the same mugshot gallery. Clearly our model still makes a significant impact on retrieval performance, despite the sketch-composite domain shift.

6. Conclusions

We investigated two questions: Whether it is possible to improve facial sketches whose quality is impacted by a large delay between seeing the face and making the sketch; and whether such models can be used to improve practical forensic sketch recognition. We were able to demonstrate that it is indeed possible to improve facial sketches drawn after a time-delay, and that this translates into the significantly improved state of the art performance on the important task of forensic sketch matching.

One limitation of our current work is that each HoG dimension is modelled independently, so cross-pixel correlation is not exploited. In future, we would explore richer information sharing architectures, such as local patches, CRF smoothing, and multi-task among neighboring pixels. Secondly, we ultimately exploited the contributions of cross-modal and communication gaps only implicitly via MTL sharing. A richer framework more explicitly modelling the contributing factors should be explored.

Acknowledgements: This project received support from the European Union's Horizon 2020 research and innovation programme under grant agreement #640891, and the Royal Society and Natural Science Foundation of China (NSFC) joint grant #IE141387 and #61511130081. We especially would like to thank the China Scholarship Council (CSC) for funding the first author to conduct the entirety of this project at Queen Mary University of London.

References

[1] H. S. Bhatt, S. Bharadwaj, R. Singh, and M. Vatsa. On matching sketches with digital face images. In *BTAS*, 2010.

[2] H. S. Bhatt, S. Bharadwaj, R. Singh, and M. Vatsa. Memetically optimized mcwld for matching sketches with digital face images. *TIFS*, 2012.

[3] E. V. Bonilla, K. M. A. Chai, and C. K. I. Williams. Multitask gaussian process prediction. In *NIPS*, 2008.

[4] J. Choi, A. Sharma, D. W. Jacobs, and L. S. Davis. Data insufficiency in sketch versus photo face recognition. In *CVPR*, 2012.

[5] C. Frowd. *Introduction to Applied Psychology*, chapter Eyewitnesses and the use and application of cognitive theory. 2011.

[6] C. Frowd, V. Bruce, A. McIntyre, and P. Hancock. The relative importance of external and internal features of facial composites. *British Journal of Psychology*, 2007.

[7] C. Frowd, W. Erickson, J. Lampinen, F. Skelton, A. McIntyre, and P. Hancock. A decade of evolving composite techniques: Regression-and meta-analysis. *Journal of Forensic Practice (in press)*, 2015.

[8] H. Galoogahi and T. Sim. Inter-modality face sketch recognition. In *ICME*, 2012.

[9] G. Hu, Y. Yang, D. Yi, J. Kittler, W. Christmas, S. Z. Li, and T. M. Hospedales. When face recognition meets with deep learning: an evaluation of convolutional neural networks for face recognition. In *ICCV Workshops ChaLearn Looking at People*, 2015.

[10] R. G. U. Jr. and N. da Victoria Lobo. A framework for recognizing a facial image from a police sketch. In *CVPR*, 1996.

[11] B. F. Klare and A. K. Jain. Heterogeneous face recognition using kernel prototype similarities. *TPAMI*, 2013.

[12] B. F. Klare, Z. Li, and A. K. Jain. Matching forensic sketches to mug shot photos. *TPAMI*, 2011.

[13] A. Kumar and H. D. III. Learning task grouping and overlap in multi-task learning. In *ICML*, 2012.

[14] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *ICCV*, 2009.

[15] Z. Lei, M. Pietikainen, and S. Z. Li. Learning discriminant face descriptor. *TPAMI*, 2014.

[16] P. Luo, X. Wang, and X. Tang. A deep sum-product architecture for robust facial attributes analysis. In *ICCV*, 2013.

[17] S. Ouyang, T. Hospedales, Y.-Z. Song, and X. Li. A survey on heterogeneous face recognition: Sketch, infra-red, 3d and low-resolution. *arXiv preprint arXiv:1409.5114*, 2014.

[18] S. Ouyang, T. M. Hospedales, Y.-Z. Song, and X. Li. Cross-modal face matching: Beyond viewed sketches. In *ACCV*, 2014.

[19] C. E. Rasmussen and C. K. I. Williams. Gaussian processes for machine learning. In *Gaussian Processes for Machine Learning*, 2006.

[20] A. Sharma and D. W. Jacobs. Bypassing synthesis pls for face recognition with pose, low-resolution and sketch. In *CVPR*, 2011.

[21] C. Vondrick, A. Khosla, T. Malisiewicz, and A. Torralba. Hoggles: Visualizing object detection features. *ICCV*, 2013.

[22] N. Wang, D. Tao, X. Gao, X. Li, and J. Li. A comprehensive survey to face hallucination. *IJCV*, 2014.

[23] X. Wang and X. Tang. Face photo-sketch synthesis and recognition. *TPAMI*, 2009.

[24] Y. Yang and T. M. Hospedales. A unified perspective on multi-domain and multi-task learning. In *ICLR*, 2015.

[25] A. W. Young, D. Hay, K. H. McWeeny, B. M. Flude, and A. W. Ellis. Matching familiar and unfamiliar faces on internal and external features. *Perception*, 1985.

[26] J. Zhang, N. Wang, X. Gao, D. Tao, and X. Li. Face sketch-photo synthesis based on support vector regression. In *ICIP*, 2011.