一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report

目撃証言にもとづく正面顔似顔絵スケッチの生成：CLISOTS

阿見　翼† 小森　政嗣†

† 大阪電気通信大学 情報通信工学部　〒572–0833 大阪府寝屋川市初町 18–8
E-mail: †gp20a004@oecu.jp, ††komori@osakac.ac.jp

あらまし　犯罪捜査の現場において，容疑者に関する目撃証言の共有は極めて重要である．目撃者たちの記憶を歪めることを避けるため，近年ではリアルなモンタージュ写真よりもスケッチ画像（似顔絵）が使用されるようになってきた．しかし，目撃証言から容疑者の顔を正確にスケッチすることは，すべての捜査員にとっては難しい課題である．言語的な証言から正面顔のスケッチ画像を生成するシステム（CLISOTS: CLIP based Semantic Oriented Testimony to Portrait Sketch）を提案する．このシステムは，目撃証言に含まれる抽象的な顔特徴の表現を，逐次的に顔スケッチ画像に反映させることを可能にしている．

キーワード　目撃証言, 似顔絵, CLIP, 意味的損失, 幾何学的損失

CLISOTS: CLIP based Semantic Oriented Testimony to Portrait Sketch

Tsubasa AMI† and Masashi KOMORI†

† Faculty of Information and Communication Engineering, Osaka Electro-Communication University
18–8 Hatsu-cho, Neyagawa-shi, 572–0833 Japan
E-mail: †gp20a004@oecu.jp, ††komori@osakac.ac.jp

Abstract　In criminal investigations, eyewitness testimonies about suspects are crucial. To avoid distorting these memories, investigators are increasingly turning to sketch images (composite sketches) rather than realistic montage photos. However, it's challenging for every investigator to accurately sketch a suspect's face from such testimonies. In this study, we introduce a system, CLISOTS (CLIP based Semantic Oriented Testimony to Portrait Sketch), which generates frontal face sketches from verbal descriptions. This system can effectively translate abstract facial features from testimonies into sketch details sequentially.

Key words　eyewitness testimony, composite sketch, CLIP, semantic loss, geometric loss

1. はじめに

警察の捜査活動では，事件を起こした疑いのある容疑者を特定し確保することが必要となる．容疑者の特定や捜索の際に，目撃証言をもとに容疑者の顔貌に関する情報が捜査関係者に共有されることはしばしば行われる．かつては，このような場面ではモンタージュ写真がしばしば使われていた．モンタージュ写真とは，目撃証言から，容疑者の特徴に近い眉，目，鼻，口などの部位を選択し，容疑者の顔写真を合成する手法である．しかしモンタージュ写真には，各部位の選択肢を見ているうちに，目撃者の記憶が歪められてしまう，部位ごとに容疑者に似ている特徴を選んでも全体としては似ていないことがある，完成した写真が具体的であるため捜査範囲が狭められてしまうといった問題があったため，今日ではほとんど使われなくなっている．[1]

モンタージュ写真に代わって，容疑者の顔貌に関する情報を共有するために用いられるのが似顔絵である．似顔絵を用いた捜査は「似顔絵捜査」と呼ばれ，目撃者の記憶を歪める可能性が低い，全体的な雰囲気を表現できる，素早く容疑者に関する情報が共有できるなどの特徴があることが知られている．一方で，目撃証言のみから捜査活動に役立つ似顔絵を描くことは，一般の捜査員にとっては容易なことではない．現在各地の都道府県の警察では，「似顔絵捜査官」や「似顔絵捜査員」を養成するために，写真や文字などの情報をもとに絵を描く実技試験を行った上で専門の講習を受ける制度が設けられているが，すべての捜査員を似顔絵捜査官に育成するのは容易ではない．

そこで，本研究では，言語的な表現から逐次的に正面顔スケッチ画像を生成するシステム（CLISOTS: CLIP based Semantic Oriented Testimony to portrait Sketch [1]）を提案する．ここで，スケッチ画像とは比較的少数の線で描かれた対象物の画像とする．

（注1）：https://github.com/tsubasa652/CLISOTS

(a) エッジ検出 [2]　(b) Informative Drawings [4]　(c) CLIPasso [5]

図 1: 画像を入力とするスケッチ画像生成手法．すべて東アジア人平均顔画像 [6] から作成された．

Fig. 1　Sketch image generation method using images as input. All were created from East Asian average face images [6].

スケッチ画像は，リアルな顔画像よりも目撃者の記憶を歪める危険性を抑制しつつ，容疑者に関する情報共有を可能にすると考えられる．言語情報のみから手軽にスケッチ画像を生成し，またそのスケッチ画像を言語情報により更新していくシステムがあれば，有効な捜査手法となるだろう．

これまでにもスケッチ画像を生成する手法はいくつか提案されてきた．一つは顔画像からスケッチ画像を生成する手法である．輝度値から物体の境目を検出するエッジ検出 [2] や，敵対的生成ネットワーク（GAN）と CLIP [3] を用いることで自然なスケッチを生成する Informative Drawings [4] などが提案されてきた（図 1）．さらに CLIPasso [5] では，入力画像の幾何学的特徴だけでなく CLIP [3] により行った画像の意味的な解釈をもとに生成された画像の評価を行うことで，抽象的なスケッチを生成することに成功している．しかしこれらの手法では言語的な情報（すなわち目撃証言）をスケッチ画像に反映させることができない．

言語情報を反映させたスケッチ画像の生成手法としては，CLIPDraw [7] という手法が提案されている．この手法では，初期値として微分可能なベジェ曲線のランダムなセットを与え，さらに生成されたベジェ曲線によるスケッチ画像と，与えられた言語的なプロンプトの双方を CLIP に投入し，これらの間の意味的損失が最小になるように最適化を行うことでスケッチ画像を生成している．しかし，CLIPDraw では意味的損失のみを考慮しており，どのような画像が生成されるかを制御することは困難である．似顔絵捜査で用いられるような正面顔スケッチ画像を生成するためには，幾何学的な損失も考慮する必要がある．

本研究では CLIPDraw [7] で提案された言語的プロンプトに基づくベジェ曲線の生成手法と，CLIPasso [5] で提案されている幾何学的性質を保持したスケッチの生成手法を組み合わせることで，正面顔スケッチ画像を言語的な記述のみから生成するシステム（CLISOTS）を提案する．

2. CLISOTS システムの概要

本システムでは，言語的に表現された顔特徴をベジェ曲線で表現するスケッチ化を行う（図 2）．顔特徴のスケッチ化には CLIPDraw [7] および CLIPasso [5] で提案された微分可能なラスタライザを用いている．また，このベジェ曲線のパラメータを言語的に表現された顔特徴に基づいて更新することで，似顔絵の生成を行う．

正面顔のスケッチを作成するには，言語情報に加えて顔形状の形態的な情報を考慮してベジェ曲線のパラメータを決定する必要がある．本システムでは，顔の形態的な情報として，初回のスケッチ作成では顔画像を参照画像とし，2 回目以降のスケッチ作成では前回に作成されたスケッチ画像を利用する．この際，顔の顕著な特徴を重点的にスケッチ化の対象とするために，予め顔の重要な標識点を設定した上で，平均顔画像から MediaPipe [8] により標識点座標を取得し，その周辺領域を着目領域としている．第 1 回目の生成に用いる画像には StyleGAN2 で生成された東アジア人の平均顔 [6] を使用した（図 3a）．ただし服が描画対象となることを避けるため，服の部分は背景色に変更した．また顔の特徴を捉えるための標識点座標に MediaPipe FaceMesh [8] の全 468 点の中から 44 点を選んだ点を用いた．また 2 つの瞳孔位置に対応する座標を算出するため，各 2 点の座標の平均値を求め，これらも特徴点座標として用いた（図 3b）．結果的に 46 点の顔特徴点を用いた．これに従いスケッチに用いられるベジェ曲線の本数は 46 本となった．

スケッチパラメータの更新は意味的損失と幾何学的損失の 2 つの損失の和をもとに行われる．意味的損失は，入力した言語的な情報，すなわち目撃証言が，描画されるスケッチにどの程度反映されているかを表す値である．意味的損失の算出には CLIP [3] を用いるが，目撃証言では曖昧で抽象的で感覚的な表現（たとえば「真面目そうな」）といった表現がなされるため，そのままでは CLIP への入力には適さない．そこで，本システムでは入力したテキストを一度 ChatGPT-4 に与えることで，抽象的な表現と関連すると考えられる具体的な顔の特徴に変換した上で，それを CLIP のプロンプトとする．このプロンプトを ViT-B/32 CLIP モデルでエンコードした出力と．上述のラスタライザの出力を同じモデルによりエンコードした出力のコサイン類似度を求め，この値を意味的損失とする．

$$L_{semantic} = dist(CLIP(I), CLIP(R(\{s_i\}_{i=1}^{n}))) \tag{1}$$

$$dist(x, y) = 1 - \frac{x \cdot y}{||x|| \cdot ||y||} \tag{2}$$

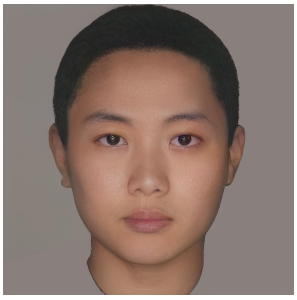
幾何学的損失は，生成されたスケッチが元画像から大幅に乖離しないために用いられるものであり，生成されたスケッチと元画像をともに ResNet101 CLIP モデルに入力した際の中間層の L2 ノルムにより表現される．元画像には初期スケッチの場合は日本人の平均顔画像を用いる．また 2 回目以降のスケッチ生成では，前回生成されたスケッチを用いる．

$$L_{geometric} = \sum_{l} ||CLIP_l(I) - CLIP_l(R(\{s_i\}_{i=1}^{n}))||_2^2 \tag{3}$$

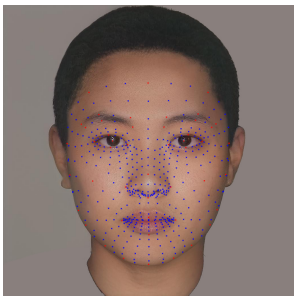
これらの意味的損失と幾何学的損失を重み付けした和を求め，これを損失関数とした．

$$\min_{\{s_i\}_{i=1}^{n}} L_{geometric} + w_s \cdot L_{semantic} \tag{4}$$

この損失関数を最小化するパラメータをバックプロパゲーションにより更新し探索することで，言語的な入力を反映させた正面顔スケッチを生成することを行った．意味的損失の重み w_s は目撃証言を反映させるため，正面顔画像が生成される範囲内で最も大きい値が望ましい．予備調査に基づき初回スケッチ生成時の意味的損失の重み w_s は，$w_s = 1.75$ とした（図 4）．また 2 回目以降の生成では，スケッチを参照画像とするため，意味的損失の重み w_s を 0.75 に設定した．また，1 回の描画に必要な最適化のための繰り返しの回数は 2000 回に設定した．



(a) 初期画像として用いた東アジア人平均顔 [6]



(b) 注目領域決定のために用いた 46 点の顔特徴点座標

図 3: 初期画像と着目点

Fig. 3　Default image and points of interest

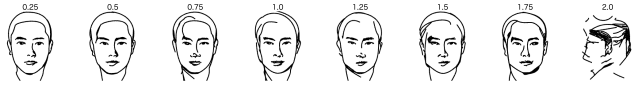


図 4: 意味的損失 w_s と生成された第 1 回目生成スケッチ画像の関係．ただし，与えたプロンプトは「30 代のアクティブな感じのするサラリーマンで髪の毛は短め」である．

Fig. 4　Relationship between semantic loss w_s and the first generated sketch image. However, the prompt given was “30 代のアクティブな感じのするサラリーマンで髪の毛は短め”.

3.　結果と考察

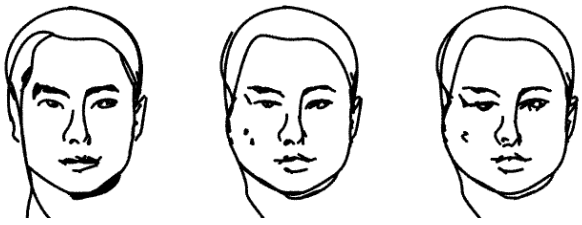
本システムを用いて，2 つのプロンプトセットに基づいて似顔絵生成を行った（表 1）．そのプロンプトセットに基づく生成画像を図 5 と図 6 に示す．各画像の生成には Google Colaboratory を使用して概ね 7 分を要した．

プロンプトセット 1 の第 1 証言「30 代のアクティブな感じのするサラリーマンで髪の毛は短め」は，“In his 30s, he emits an aura brimming with energy, his facial features conveying a sense of toughness. His well-groomed short hair, not over-extended, accentuates his business-like ambiance. ”，第 2 証言「もっと太っている」は “They have a round face with plump cheeks.” に，第 3 証言「だんご鼻」は “The feature of having a round, short nose, and an overall plump shape.” と ChatGPT により CLIP 向けに翻訳された．それぞれ生成されたスケッチ画像には，与られた各証言が反映されていたと言える．

プロンプトセット 2 は，顔画像（図 6）を実験参加者に見せ，この顔特徴を口頭で表現させることにより作成した．この顔画像は，我々の先行研究 [6] で生成された「最も信頼性が低い顔」の画像であり，実在する人物ではない．第 1 証言は “A man in his mid-30s, characterized by thick eyebrows. He has short hair and sports a rare-toned blueish beard. His gaze is somewhat intimidating, giving off a frightening impression at first glance.”，第 2 証言は “High cheekbones and a long, narrow chin.”，第 3 証言は “They have the characteristic of having sunken cheeks, which makes their face appear long and slender.” と翻訳された．第 1 証言は ChatGPT により誤訳されており，結果的にスケッチ画像に髪が描かれてしまっていることがわかる．また第 2 回と第 3 回の生成画像に大きな違いがなく，第 3 証言が画像生成に十分には反映されていなかった．

表 1: CLISOTS に与えたプロンプト
Table 1　Prompts given to CLISOTS

	プロンプトセット 1	プロンプトセット 2
第 1 証言	30 代のアクティブな感じのするサラリーマンで髪の毛は短め	30 代で眉毛が太くて、青髭の短髪で目つきが悪い
第 2 証言	もっと太っている	もう少し顔が縦長
第 3 証言	だんご鼻だった	頬がこけている



(a) 第 1 証言　(b) 第 2 証言　(c) 第 3 証言

図 5: プロンプトセット 1 によって生成されたスケッチ画像

Fig. 5　Sketch image generated by Prompt Set 1.

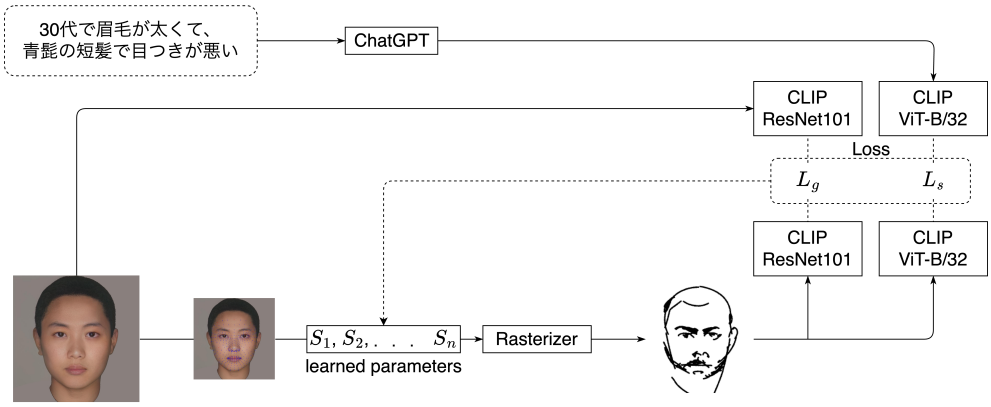
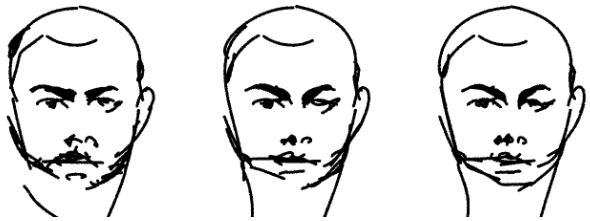


図 2: 本システムの概要

Fig. 2　Overview of this system.

図 6: 目撃証言作成のために参照した画像 [6]

Fig. 6　Images referenced to create eyewitness testimony [6]



(a) 第 1 証言　(b) 第 2 証言　(c) 第 3 証言

図 7: プロンプトセット 2 によって生成されたスケッチ画像

Fig. 7　Sketch image generated by Prompt Set 2.

4. 結　　論

本研究では，ベジェ曲線を用いたスケッチ生成手法 CLIPasso [5] を参考に，目撃証言のみから人物の正面顔のスケッチ画像を生成するシステム（CLISOTS）を提案した．本システムの特徴は以下の 4 点にまとめられる；(1) 目撃証言を反映するために生成されたスケッチ画像と目撃証言の間の意味的損失を導入したことで，言語的な記述をスケッチに反映させることが可能となった，(2) GPT-4 により曖昧な言語的表現をより具体的な顔特徴を表現したプロンプトとして CLIP に投入することができる，(3) 東アジア人平均顔を参照画像とし，予め顔の特徴で注目すべき点を指定していることで，全体的な顔特徴を余すところなく生成画像に反映することができるようになった，(4) 生成されたスケッチ画像を参照画像とすることで，逐次的な変更が可能となった．本システムは，手軽に人物のスケッチ画像を作成できるという利点があるが，生成時間が長いという問題がある．今後，より短時間でスケッチ画像を生成できるように改良を加えることで，目撃証言をもとにした似顔絵捜査がより容易になることが期待できる．

謝辞

本研究は JSPS 科研費 23K03021 の助成を受けた．

文　　献

[1] 渡邊伸行. 捜査用似顔絵描画の実際. 日本顔学会誌, Vol. 22, No. 2, pp. 61–67, 2022.

[2] John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. PAMI-8, No. 6, pp. 679–698, 1986.

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

[4] Caroline Chan, Frédo Durand, and Phillip Isola. Learning to generate line drawings that convey geometry and semantics. In *CVPR*, 2022.

[5] Yael Vinker, Ehsan Pajouheshgar, Jessica Y Bo, Roman Christian Bachmann, Amit Haim Bermano, Daniel Cohen-Or, Amir Zamir, and Ariel Shamir. Clipasso: Semantically-aware object sketching. *ACM Transactions on Graphics (TOG)*, Vol. 41, No. 4, pp. 1–11, 2022.

[6] Keito Shiroshita, Masashi Komori, Koyo Nakamura, Maiko Kobayashi, and Katsumi Watanabe. Application of gaussian process preference learning for visualizing facial features related to personality traits. In *2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, pp. 1–6, 2021.

[7] Kevin Frans, Lisa Soros, and Olaf Witkowski. Clipdraw: Exploring text-to-drawing synthesis through language-image encoders. *Advances in Neural Information Processing Systems*, Vol. 35, pp. 5207–5218, 2022.

[8] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, et al. Mediapipe: A framework for perceiving and processing reality. In *Third workshop on computer vision for AR/VR at IEEE computer vision and pattern recognition (CVPR)*, Vol. 2019, 2019.