Vol.2018-HCI-180 No.20
Vol.2018-UBI-60 No.20
2018/12/5

制約つきグラフ描画によるユーザインタフェースを用いたダイジェスト動画作成

山下 紗季[1,a)]　伊藤 貴之[1,b)]　Tobias Czauderna[2,c)]　Michael Wybrow[2,d)]

概要：著者らは映像内に登場する特定人物に注目してダイジェスト映像を生成する一手法を提案している．本手法ではショット選択のためのユーザインタフェースを生成し，その上で顔識別結果にもとづいて自動選択されたショットとユーザによって選択されたショットを連結する．これにより人物が自動検出されなかったショットもダイジェスト動画の一部に選ぶことができ，より満足度の高いダイジェスト動画を生成できる．本報告ではその一環として，人物が自動検出されなかったショットを選出するためのユーザインタフェースを提案する．ユーザインタフェースの生成には制約つきのグラフ描画を用いる．力学モデルに制約を加えることでサムネイルの配置を調整する．本手法の主な用途として，俳優やミュージシャンなど人物の魅力的なカットを集めて楽しむという用途があげられる．

Digest Movie Creation with a User Interface Applying Constraint-Based Graph Layout

Saki Yamashita[1,a)]　Takayuki Itoh[1,b)]　Tobias Czauderna[2,c)]　Michael Wybrow[2,d)]

Abstract: We are developing a new method to generate digest videos focusing on specific persons appearing in the video. This method generates a user interface for shot selection. We suppose to manually select shots on the user interface and then combine them with automatically selected shots to generate a digest video. As a result, we can insert the shots in which the target is not automatically detected as part of the digest videos, and generate highly satisfied digest videos. This paper presents a user interface which assists the manual selection of preferable shots. We apply a constraint-based graph layout for generating the user interface. The system adjusts the arrangement of the thumbnails by applying constraints to the force-directed graph layout. This method aims to collect attractive scenes of specific persons such as actors or musicians.

1. はじめに

ダイジェスト動画は長時間の動画コレクションの中から必要なシーンだけを短時間で鑑賞する有効な手段である．本報告では，映像内に登場する人物に注目したダイジェスト動画生成を支援する一手法を提案する．

本研究におけるダイジェスト動画の定義は，与えられた動画群からユーザが指定した人物が映るシーンを検出して連結させたものである．このようなダイジェスト動画が生成されることで，グループ歌手の映像やドラマ映像からユーザが鑑賞したい人物にのみ注目した短い動画を生成することができる．特にユーザがグループ内の特定の個人や特定の俳優のファンである場合に，このようなダイジェスト動画は有用である．なお，本研究では動画の内容要約は目的としない．

著者らは映像内に登場する特定人物に注目してダイジェスト映像を生成する一手法を提案している．本手法では入力映像を多数のショットに分割し，その各々に顔画像認識を適用する．その結果として，特定人物が確実に含まれると判定されたショットを「正解ショット」とし，ダイジェスト映像を構成するショットとする．さらに，正解ショッ

[1] お茶の水女子大学　Ochanomizu University
[2] Monash University
[a)] shanxia@itolab.is.ocha.ac.jp
[b)] itot@is.ocha.ac.jp
[c)] tobias.czauderna@monash.edu
[d)] michael.wybrow@monash.edu

トのいずれかに対して一定以上の類似度を有するショットを「候補ショット」とする．候補ショットは指定された人物を含む可能性はあるが確定的ではないショットと考えることができる．本手法ではユーザインタフェースを介して候補ショットをユーザに提示し，ダイジェスト映像に含まれるべきショットを選択させる．動画像処理によって自動選択された正解ショットとユーザによって選択された候補ショットを組み合わせることで，少ない操作で満足度の高いダイジェスト動画を生成する．

本報告ではその一環として，時間的に隣接するショットの接続関係を表示し，生成されるダイジェスト動画を概観しながらショットを選択できるユーザインタフェースを提案する．ユーザインタフェースの生成には制約つきのグラフ描画を用いており，力学モデルに制約を加えることでサムネイルの配置を調整する．

2. 関連研究

ビデオから特定人物を検出する手法として，まず Chen らの手法 [1] があげられる．この手法は対象となる入力ビデオを報道番組に限定した上で，顔識別で得られる情報のほかにテキスト情報，タイミング情報なども参照して特定人物を検出する．報道番組以外のビデオ（例えば音楽映像やドラマ映像）に適用する場合には，別の手法を併用する必要がある．また，平井ら [2] は顔に特化した認証手法を提案し，ミュージックビデオを対象とした実験で個人アーティストに対して 95%の認証率を実現している．しかし，顔がカメラを向いていないショットや，手元など顔以外の部分にクローズアップしているショットなどは顔領域が検出されず，顔認証ができない．そのため，ユーザが指定した人物が映るショットすべてを検出することは難しい．

動画編集を支援するためにユーザインタフェースを生成する手法も多数報告されている．一例としてここでは土田らの手法 [3] をあげる．このシステムでは，複数のカメラから同時に撮影されたダンス映像を自動編集するだけでなく，ユーザインタフェースを生成することで好みのダンス映像に調整できる仕組みをユーザに提供する．この手法でのビデオ自動編集は，土田らが調査した動画編集の原則に基づいている．一方でこの手法は，ダンス映像に特化している点，および多視点で同時録画された映像を題材にしている点において本手法とは前提条件が異なる．

多数のコンテンツを連結させるユーザインタフェースは，ダイジェスト動画生成に限定しなければ多数報告されている．一例として後藤ら [4] は，インタラクティブに選曲しプレイリストを構築するための音楽再生用のユーザインタフェースを提案した．このシステムは類似する楽曲を自動的に連結する機能を有している．ユーザが画面上でアイテムを並べてシーケンスを構築できるという点では本手法と類似しているが，後藤らのシステムは 1 つのプレイリストの作成が主目的ではないという点で本手法と異なる．

3. ユーザインタフェースの設計

本報告では図 1 のようなユーザインタフェースを提案する．画面の 4 辺にダイジェスト動画に採用することが決定されたショット（以下「採用済みショット」と呼ぶ）を連結して配置し，その内側に採用候補となるショットを配置する．ユーザは採用したいショットを採用済みショット列の任意の位置へドラッグすることにより，ショットの採用と挿入ができる．候補となるショットから採用済みショットへ接続されている線分は，システムが推薦する挿入位置を表す．これによってユーザの挿入操作を支援する．

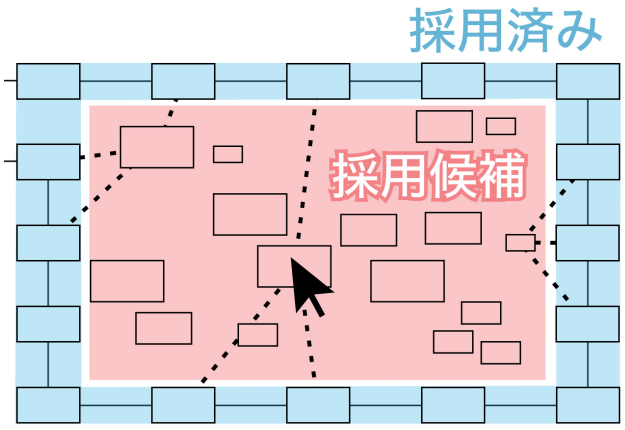

図 1　ユーザインタフェースの設計

4. ユーザインタフェース生成の前処理

本章では，指定された人物の自動的な検出や，ユーザインタフェース生成のための情報を取得する処理について述べる．

4.1 ショット分割

まず，入力された動画をショットに分割する．ショットとは，場面が大きく変化するカット点に挟まれた連続したフレームを指す．このショットが，生成されるダイジェスト動画の一単位となる．分割処理には Panagiotis ら [5] のプログラムを用いた．このプログラムからは各ショットの始点と終点をフレーム番号で取得できる．

4.2 顔検出と顔識別にもとづく得点付与

続いて各ショット中の顔領域から指定人物を含む可能性を推定し，ショットに得点を与える．

はじめにショット中の顔領域を検出する．顔領域検出には Microsodt Azure[6] の Media Services を用いた．顔検出できたショットについては，ユーザに指定人物の顔画像を入力させ検出された顔領域との類似度を範囲 $[0.0, 1.0]$ の実数で算出し，その類似度を得点とする．類似度は Microsodt

Azure の Face API を用いて算出する．顔領域が検出されなかったショットについては，顔検出されたショットの得点をもとに得点を算出する．得点を求めたいショット A の得点を P_A として，ショット A と顔検出できたショット群 B_i との類似度をそれぞれ求める．類似度を $Sim(A, B_i)$ としたときに，以下の式で表される実数をショット A の得点を P_A とする．

$$P_A = max(\ P_{B_i} Sim(A, B_i)\ ) \tag{1}$$

これにより顔領域の条件の差を吸収したショット選出を可能にする．類似度の判定には AKAZE 特徴量 [7] を用いる．

そして $[0, 1]$ の間に閾値を 2 つ定め，それらを $s, t\ (s < t)$ としたとき，$P_A < s$ となるショットは指定人物が存在しないであろうとしてダイジェスト動画に組み込むショットの候補から除外する．ここで，$P_A > t$ となるショットは確実に指定人物を含んでいるとして，あらかじめダイジェスト動画に採用する．これを「正解ショット」と呼ぶ．一方，$s \leq P_A \leq t$ であるショットは指定人物を含む可能性はあるが確定的ではないとし，ユーザによる選択でダイジェスト動画に組み込む．これを「候補ショット」と呼ぶ．

4.3 特徴量算出

4.2 節で取得した得点のほかに，各ショットから特徴量を算出する．現状で算出している特徴量は，ショットの長さ，入力動画における時間上の位置，顔の大きさ，顔の位置，指定人物以外の顔の数，画面の動き方向である．これらの特徴量は，ユーザによるショット選択時にどのような内容のダイジェスト動画にするかを考慮するための指標として用いられる．ショットの長さと時間上の位置は 4.1 節で取得したカット点情報から算出される．顔に関する特徴量は顔検出の結果から算出される．画面の動き方向はオプティカルフローをもとに算出される．

4.4 候補ショットの挿入位置の推薦

続いて候補ショットの各々について，どの正解ショットの前後（これを挿入位置と呼ぶ）に挿入するのがふさわしいかを判定する．ここで，ある候補ショットを C，正解ショット群を D_i とし，挿入の推薦度を $R(C, D_i)$ とする．CD_i 間の画像的類似度を Im，入力動画における時間的近接度を Tm としたとき，以下の式によって $R(C, D_i)$ を求める．

$$R(C, D_i) = Tm + (1 - Tm)Im \tag{2}$$

$$Tm = (Dist(C, D_i) - 1)^6 \tag{3}$$

$Dist(C, D_i)$ は C と D_i のショット番号の差を求め，その絶対値を全体のショット数で正規化したものである．図 2 に示すように，$Dist$ が 0 に近づくと時間的近接度 Tm が大きくなる．これによって画像内容が類似している正解ショットの前後のみならず，時系列的に近接している正解ショットの前後も挿入位置として推薦することができる．

以上の方法により各候補ショットについて $R(C, D_i)$ を算出し，その値が 1〜3 位となる正解ショット D_i を本手法が推薦する挿入位置とする．

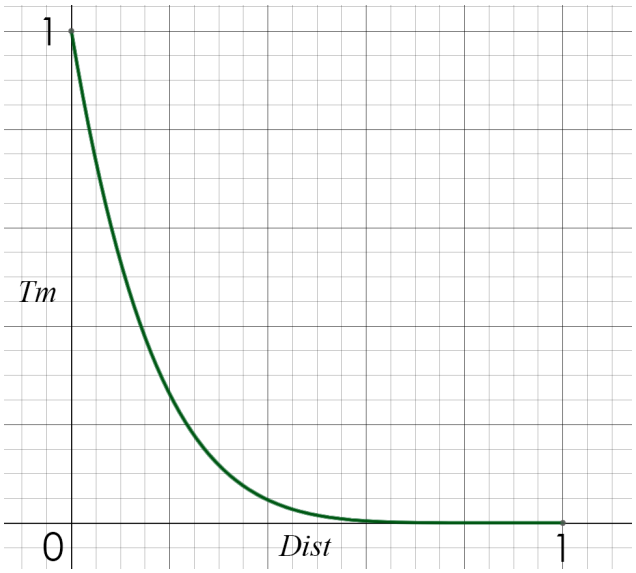

図 2 ショット番号の差 $Dist$ と時間的近接度 Tm の関係

5. ユーザインタフェースの生成

本章では，ショット連結順を確認しながらショット選択ができるユーザインタフェースを提案する．ユーザインタフェースは，各ショットのサムネイルをノードとしたグラフとして生成される．グラフのエッジは，時間的に隣接する 2 つの採用済みショットの連結，および候補ショットに対してシステムが推薦する挿入位置への連結を表現する．ここで採用済みショットは，自動的にダイジェスト動画に採用された正解ショット，およびユーザ操作によって選択された候補ショットで構成される．

5.1 採用済みショットの配置

4.2 節で説明したとおり，得点が $P_A > t$ となるショットは正解ショットとしてあらかじめダイジェスト動画に採用される．ユーザインタフェースの生成に際してまず，この正解ショットのみを対象として仮の再生順を決定し連結する．そして図 1 のように，正解ショット群を画面の 4 辺に連結して表示する．正解ショットの表示には，一辺の上に位置を固定する制約を加える．

5.2 候補ショットの配置

続いて仮連結された正解ショットの内側に候補ショットを配置する．候補ショットの画面上の位置の算出には力学

指向ノード配置手法を用いる．このとき各候補ショットを正解ショットよりも外側に配置しない制約と，ノードを重ねて表示しない制約を加える．候補ショットとシステムが推薦する挿入位置はエッジで接続されているため，ユーザは各候補ショットの位置から挿入位置を推察することができる．サムネイルの大きさは 4.2 節で求めた得点に比例した大きさとする．あるいは，4.3 節の特徴量のうち一つを選択し，その値に比例した大きさとする．

5.3 ショットの挿入と削除

ユーザは候補ショットを正解ショット列の任意の位置にドラッグすることでショットを挿入する．マウスカーソルを候補ショットの上に移動させると，挿入位置となる正解ショットへのエッジが表示される．またマウスボタンを押下すると，その間サムネイルを拡大して表示する．以上によりユーザは，挿入位置となるショットについて，サムネイルやダイジェスト動画における時間的な位置，前後に仮接続されたショットを確認することができる．これらの情報をもとに，マウスカーソルを置いているショットを採用するか否か，および挿入位置を選択する．ショットの挿入が行われると，ユーザインタフェースは採用済みのショット列を更新し 5.1 節と同様の処理によって再配置する．

また，候補ショットをダブルクリックするとそのショットを削除する．不要なショットを消去することで候補ショットを絞り込むことができ，ユーザの選択操作を支援する．

6. ユーザインタフェースの実行例

本章ではユーザインタフェースの実行例を紹介する．入力動画として 1 本のミュージックビデオを使用し，4.1 節の処理によって 145 個のショットに分割された．この例では得点が 0.8 より大きいショットを正解ショットとしてあらかじめ採用し，得点が 0.05 より小さいショットは候補ショットにも該当しないとして表示対象から削除した．正解ショットの連結順序は入力動画における時系列順とした．また，現在 4.4 節の画像的類似度 Im として 4.2 節で算出した AKAZE 特徴量を使用した．ユーザインタフェースの実装には cola.js[8] を適用した．

6.1 ショットの表示

プログラムを実行すると，まず図 3 のような画面が表示される．マウスカーソルをサムネイルの上に移動させると，挿入位置の候補となるショットへのエッジが表示される．また，マウスボタンを押下すると図 4 のようにサムネイルが拡大される．ユーザが自由にショットをドラッグできるよう，マウスボタンが押されている間はノードの重なりを回避する制約を無効にする．

6.2 ダイジェスト動画の作成例

本節では第一著者が実際にユーザとなってダイジェスト動画を作成した手順を紹介する．

図 3 は本システムが生成したユーザインタフェースの初期状態を示している．この時点では，人物を含まない自動車のみのショットが候補ショットに見受けられる．そのためユーザはまず，図 5 に示すように，人物を含まない候補ショットを削除した．そして画面を概観し，指定人物についてバラエティに富んだシーンをダイジェスト動画に採用したいと考えた．ここで，指定人物の登場シーンは主に歌唱シーンと運転シーンからなり，運転シーンの方がショット数が少ない．そのためユーザは，運転シーンを優先して採用することにした．ショットを挿入する際には，図 6 に示すように，挿入したいショットを採用済みショット列へドラッグする．このときユーザインタフェースは，図 7 に示すように，マウスボタンを離したときのノードの位置から挿入位置を判定し，採用済みショット列を更新する．続いてユーザは，残りの候補ショットについても，採用済みショットに類似のものがない候補ショットを優先して採用した．このときユーザは，システムが推薦する挿入位置の中から候補ショットの挿入位置を選択した．

上記の手順で採用されたショットの列を図 8 に示す．入力動画における指定人物の登場ショットは全部で 50 ショットあるが，ユーザはそのうちの 36 ショットを採用した．この 36 ショットのうち，手動で採用したショットは 13

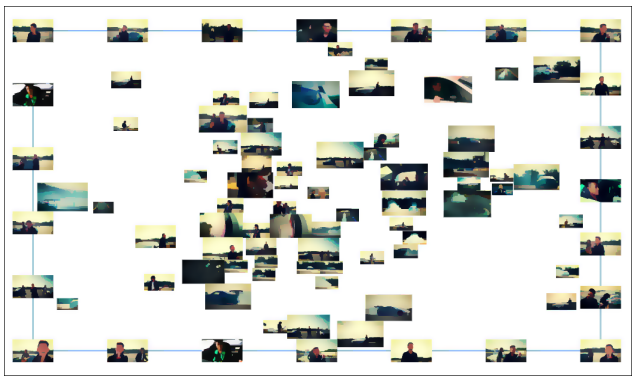

図 3 ユーザインタフェースの初期画面

図 4 クリック時

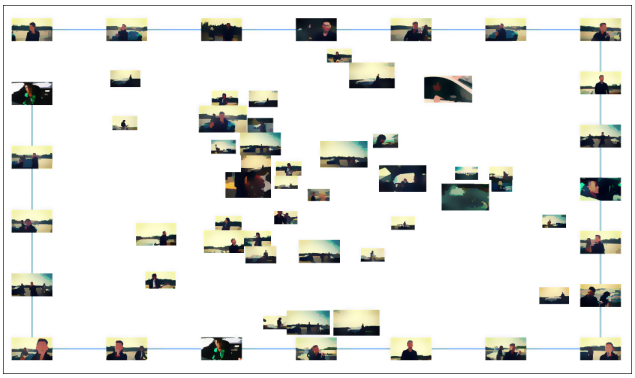

図 5　人物を含まないショットを削除したところ

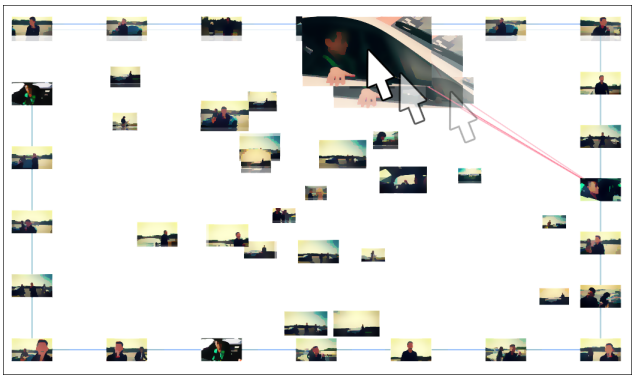

図 6　ショットの挿入

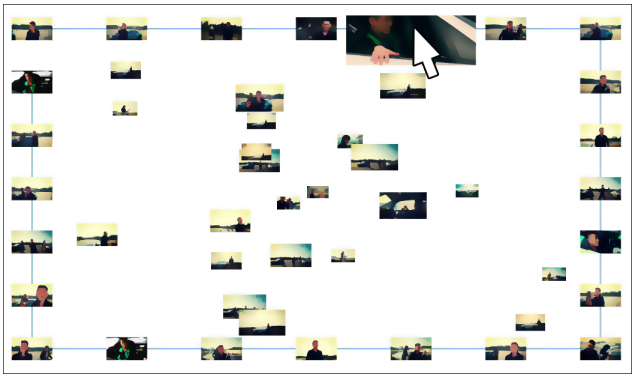

図 7　挿入されたショットを拡大して確認したところ

ショットである．歌唱シーンのバストアップが多く採用されているが，これは指定人物のバストアップが 4.2 節の顔識別処理によって高い得点となり，自動的に正解ショットとして採用されたためと考えられる．

また，6.1 節の表示時点で除外されてしまったが指定人物を含むショット群を図 9 に示す．ショットの数は全体の 6%ほどだが，指定人物の手のクローズアップや全身を遠くから撮影したショットが含まれており，このようなショットは作成したダイジェスト動画に採用されていない．

7.　まとめと今後の課題

本報告では，特定人物に注目したダイジェスト動画生成を支援する手法の研究の一環として，自動判別とユーザによる選択を組み合わせたショットを選出するためのユーザインタフェースを提案した．本手法ではユーザインタフェースの生成に制約つきグラフ描画を採用することで，サムネイルの位置調整を実現した．また本報告では，ユーザインタフェースの実行例と制作したダイジェスト動画を紹介した．

図 8　実際に作成したダイジェスト動画

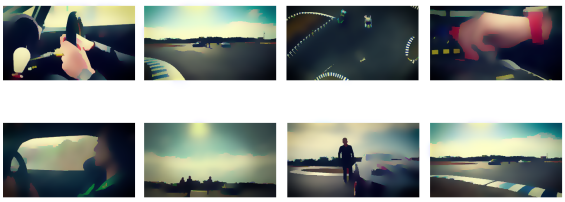

図 9　除外されてしまった指定人物のショット

今後の課題としては，まずユーザインタフェースの拡充があげられる．たとえば，候補ショットをクラスタリングして表示できるようにしたい．現在の実装ではすべての候補ショットを個別に表示しているため，ショット数の多い入力動画を使用した場合に画面が煩雑になる．類似する候補ショットをクラスタとして表示することで画面上のサムネイル数を減らせるほか，クラスタごと消去できることでユーザの手間を省いたり，クラスタから厳選したショットを採用できることでダイジェスト動画の満足度を上昇させたりといった効果が期待できる．

得点や特徴量の算出手法の改善も今後の課題としてあげられる．たとえば 6 章で記述したように，図 3 において自動車のみのショットが大きく表示されているのが見受けられる．これは 4.2 節で述べた AKAZE 特徴量を用いた類似度の算出において，自動車などエッジを多く含む画像で特徴点が多数検出され，偶発的にマッチング率が高くなるために起きていると考えられる．また，図 9 に示した手元

のクローズアップシーンのように，特定人物を写しているショットの得点が低くなる場合があり，これも状況によっては問題となりうる．このような不適切な得点を改善する手法として，一般物体認識を用いてショットに人物が含まれるかどうかを判定し，その結果を得点に反映することがあげられる．また候補ショットの挿入位置についても，現在は正解ショットのいずれかを推薦するのみで前後どちらに挿入するべきかは考慮していない．この問題を解決するために，たとえばショット A の最終フレームとショット B の先頭フレーム間で推薦度を算出する，といった処理を追加する必要がある．

そのほかの課題としては，ユーザインタフェースの操作性やダイジェスト動画の生成結果に対する評価手法の検討があげられる．

将来的には，ショットの切れ目に音声処理を施すことやユーザが指定した長さでダイジェスト動画を生成する機能の実装も検討したい．

参考文献

[1] Ming-yu Chen and Hauptmann Alexander, Searching for a specific person in broadcast news video, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'04), vol. 3, pp. iii–1036, 2004.
[2] 平井辰典, 中野倫靖, 後藤真孝, 森島繁生, シーンの連続性と顔類似度に基づく動画コンテンツ中の同一人物登場シーンの同定, 映像情報メディア学会誌, vol. 66, no. 7, pp. J251–J259, 2012.
[3] 土田修平, 深山覚, 後藤真孝, 多視点ダンス映像のインタラクティブ編集システム, 第 25 回インタラクティブシステムとソフトウェアに関するワークショップ (WISS 2017) pp. 41–46, 2017.
[4] Masataka Goto and Takayuki Goto, Musicream: New Music Playback Interface for Streaming, Sticking, Sorting, and Recalling Musical Pieces, Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR 2005), pp.404–411, 2005.
[5] Sidiropoulos Panagiotis, Mezaris Vasileios, Kompatsiaris Ioannis and Kittler Josef, Differential edit distance: A metric for scene segmentation evaluation, IEEE Transactions on Circuits and Systems for Video Technology, vol. 22, no. 6, pp. 904–914, 2012.
[6] Microsoft :
Microsoft Azure Cloud Computing Platform & Services 2018/11/7 確認 https://azure.microsoft.com/ja-jp/
[7] Pablo F. Alcantarilla, Jess Nuevo and Adrien Bartoli, Fast Explicit Diffusion for Accelerated Features in Nonlinear Scale Spaces, British Machine Vision Conference (BMVC), 2013.
[8] Tim Dwyer :
cola.js: Constraint-based Layout in the Browser 2017/12/22 確認 https://ialab.it.monash.edu/webcola/