AudiblePhoto
—周囲の環境音声を付加した写真を用いた新しい感覚のデジタル・イメージング・エクスピリアンス)—

1. 背景

本プロジェクトの目的は環境音声を付加することで，豊かな感情を持ったデジタル写真を構成することである．デジタル写真に貼付けられた環境音声は雰囲気の再現に役立ち，新しい感覚のエモーショナル・エクスピリエンスを引き出せる．また，デジタル写真の形式を劇的に再定義することで，この分野において新しいフィールドを開拓することが期待出来る．写真と音声の組み合わせは，拡張された Exchangeable Image File Format (EXIF)タグや，MP3 の ID3 タグを採用した．

本プロジェクトはスマートフォン（iPhone 3G/3GS）を基本プラットフォームとして開発を進めた．まずは，写真・音声キャプチャー・アプリケーションを開発し，そしてオーディオ風景をベースにした写真の自動タギングを実装した．その為に，周囲音声の特徴点を抽出するツール(アルゴリズム)について考察を行い,実装を行った．得られた特徴点は高次元ベクターのため，レファレンスデータとのマッチングを行う際、処理は非常に長時間を要してしまう．そこで，本プロジェクトでは高速近傍探索アルゴリズムを提案した．その結果，従来の近似最近傍探索アルゴリズムより 120 倍以上の高速化が得られ，かつレファレンス・データベースの拡大においてのスケラビリティーが良くなった．

2. 目的

- 環境音声をデジタル静止画に付加する -

本プロジェクトの基本アイデアは，カメラのシャッタを押す直前と直後の環境音声を録音し，デジタル静止画に付加することだ．実現方法としては，既存の写真のデータ構造である Exchangeable Image File Format (EXIF)を拡張し，環境音声のバイナリをユーザ定義のタグに挿入する．Flickr や Picasa のような写真共有サイトは既に EXIF タグをサポートしており，EXIF タグの解読 API も公開されている．これらは，本プロジェクトの基盤として利用出来るものである．

- 環境音声の特徴点を用いたシーン判定や自動タギング -

オーディオ解析によるオーディオの特徴点抽出やシーン判定[2]はデジタル写真の自動タギングに流用することは未だ未踏の分野である．実現出来れば，写真のタギングはかなり手軽になる．最近のユーザスタディーにより，共有された写真を実際にタグ付けしたユーザは 61 パーセントに満たない[3]．アサインされたタグを用いて，環境音声をベースにした写真のグルーピングやスライドショーも可能になる．

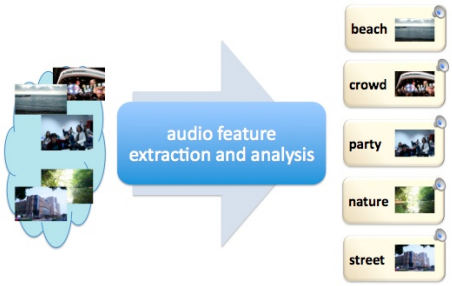

図 1. 環境音声をベースにした写真の自動タギングのイメージ.

– 緯度経度（geo-tag）や Wi-Fi のフィンガープリントの追加タグの可能性 –

本プロジェクトでは通常の GPS データ（緯度経度）に加え，デジタルコンパスデータや Wi-Fi フィンガープリントを写真に付加することによってロケーション・アウェアな写真の自動タギングが可能になる．Wi-Fi のアクセスポイントを用いた場所特定システム（PlaceEngine[4][5]）を利用すれば高精度な場所情報が得られる．行動パターンをベースにした認証技術[6]の可能性も十分実現可能である．

デジタル情報を組み込まれた写真の可能性は広いが，未だ未踏である．Flickr や Picasa といった写真共有サイトは既に EXIF タグをサポートしている．本プロジェクトのプラットフォームとしても強い基盤になる．

3. 開発の内容

AudiblePhoto project contributes and achieved results as stated below:

3.1 写真・周囲音声キャプチャー・アプリケーション(Photo and Ambient Sound Capture Application)

→ 写真と環境音声のキャプチャー・アプリケーションを iPhone3G/3GS プラットフォーム用に開発した.

→ 環境音声をカメラのシャッターが閉める直前と直後に録音された．これらは，iPhone の“タッチ”と“リリース”のイベントで制御出来る.

→ EXIF と ID3 タグを用いたデータ構造のアプローチにおいての試作を行った.

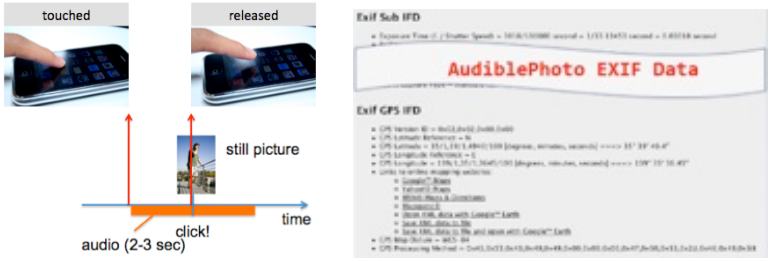

図 2. 開発プラットフォーム(左)，と EXIF タグにおいての AudiblePhoto の EXIF データの組込み.

→ iPhone 側のアプリケーション（クライアント）からサーバへ REST 接続を，Computational Auditory Scene Analysis (CASA)処理を行う.

→ AudiblePhoto ファイルを PC 側に転送する際，iPhone アプリケーション内の設定でウェブサーバを構築する.

3.2 周囲音声の特徴点抽出アルゴリズム（Ambient Sound Feature Point Extraction Solution ）

試作段階で利用したアプローチは下記の通りである：

1. The Echo Nest web API
   Echo Nest API[9]を利用し，音声の特徴点を抽出した.
2. Open Source Large Vocabulary CSR Engine (Jul ius)
   JULIUS[11]は、人声の認識を行う際に利用するツールである. JULIUS の特徴点抽出機能を利用することで、環境音声の特徴点も抽出出来るようになる.

3.3 高速近傍探索アルゴリズム（Fast Approximate Nearest Neighbor Algorithm for High -Dmensional Vector Matching Solution ）

FLANN[10]を探索のエンジンとして採用し，高速近似最近傍探索のアルゴリズムを開発した. 128 次元の 1350 個の特徴点において実験を行った結果，従来の近似最近傍探索よりも 120 倍以上の高速化が得られた(Core 2 Duo 2,4GHz 2GB システムメモリ).

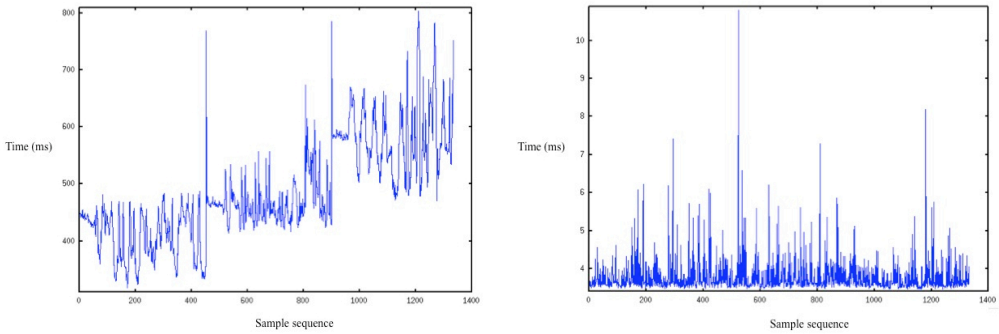

図 3. 高次元ベクターの近似最近傍探索においての実験結果. (a) 近似最近傍探索(ANN): 平均マッチング時間 = 485.38ms, (b)提案した高速近似最近傍探索: 平均マッチング時間 = 3.78ms.

3.4 周囲音声をベースにした写真の自動タギング（Ambient Sound Based Digital Photo Auto -Tagging）

Computational Auditory Scene Analysis (CASA)処理によって環境音声のシーン判定を行い，写真の自動タギングを実装した.

3.5 周囲音声タグをベースにした写真のグルーピングとビジュアリザーション（Ambient Sound Tag Based Photo Grouping and Visualization ）

本プロジェクトで開発したシステムにおいては，クライアント・サーバのアプローチを採用した. キャプチャー・アプリケーションを iPhone 側に，リソースのかかる CASA 処理はサーバ上で行う.

図4. 開発したシステムのモデル図. 写真と環境音声をiPhone3G/3GS上でキャプチャーし, 環境音声のみをCASAサーバに送信. そして, サーバの解析結果をiPhone側に送り返し, iPhone上でタグを登録する. AudiblePhotoの再生・表示は従来のMP3プレーヤー上で行える.

4. 従来の技術(または機能)との相違

本プロジェクトにおいて達成した開発成果の主な特徴は下記の通りである：

→環境音声等のデジタル情報をデジタル静止画に組込み, 用途の可能性を探究した.
→環境音声の組込み方法や環境音声のシーン判定についての開発を行った.
→高速近傍探索アルゴリズムを開発し, 従来の近似最近傍探索アルゴリズムより120倍以上の高速化を実験の結果で示した.
→写真・環境音声のキャプチャー・アプリケーションをiPhone3G/3GSプラットフォーム用に開発した.
→AudiblePhotoの写真や環境音声のビジュアリゼーションとして, 従来のMP3プレーヤー（iTunesやWinAmp等）はそのまま利用出来る.
→iPodやZune等といった携帯音楽プレーヤーにもAudiblePhotoが利用可能になっており, モバイル環境においての使用が可能になる.

5. 期待される効果

豊かな感情を持ったデジタル写真を構成することである. デジタル写真に貼付けられた環境音声は雰囲気の再現に役立ち, 新しい感覚のエモーショナル・エクスピリエンスを引き出せる. また, デジタル写真の形式を劇的に再定義することで, この分野において新しいフィールドを開拓することが期待出来る.

6. 普及(または活用)の見通し

Picasa やFlickr 等といった写真共有サービスは既にEXIFタグをサポートする. これは, 本プロジェクトにとって強力な基盤プラットフォームになることは間違いないであろう. GPSタグやデジタルコンパス等をEXIFタグに埋め込むことによってデジタル写真の新しいビジュアリゼーションの可能性を検討した.

7. 開発者名(所属)

アディヤン　ムジビヤ （東京大学大学院　学際情報学府　学際情報学専攻）

(参考)開発者URL
http://lab.rekimoto.org/members-2/adiyanmujibiya/

参考文献

[1] C. H. Chen, M. F. Weng, S. K. Jeng, and Y. Y. Chuang, “Emotion-based music visualization using photos”, in Proceedings of The 14th International Multimedia Modeling Conference (MMM 2008), Kyoto, Japan, January 2008, submitted for publication.

[2] Z. Liu, Y. Wang, and T. Chen, “Audio feature extraction and analysis for scene segmentation and classification,” *J. VLSI Signal Process. Syst.*, June 1998.

[3] Morgan Ames , Mor Naaman, Why we tag: motivations for annotation in mobile and online media, Proceedings of the SIGCHI conference on Human factors in computing systems, April 28-May 03, 2007, San Jose, California, USA

[4] Jun Rekimoto, Atsushi Shionozaki, Takahiko Sueyoshi, Takashi Miyaki, “PlaceEngine: a WiFi location platform based on realworld folksonomy”, Internet Conference 2006, pp.95-104, 2006.

[5] PlaceEngine Home Page, http://www.placeengine.com/

[6] Nobuyuki Kasuya, Takashi Miyaki, and Jun Rekimoto, “Activity-based Authentication by Ambient Wi-Fi Fingerprint Sensing”, Ubicomp 2008 Adjunct Programs, pp.16-17, 2008.

[7] iPhone-exif home page, http://code.google.com/p/iphone-exif/

[8] CLAM home page, http://www.clam.iua.upf.edu/

[9] The Echo Nest Web API, http://developer.echonest.com/

[10] Marius Muja and David G. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In International Conference on Computer Vision Theory and Applications (VISAPP’09), 2009.

[11] Open Source Large Vocabulary CSR Engine (JULIUS), http://julius.sourceforge.jp/

[12] The RESPITE CASA Toolkit Project, http://www.dcs.shef.ac.uk/spandh/projects/respite/ctk/index.html

[13] ALIPR : Automatic photo tagging and visual image search, http://alipr.com/