3C3-7

体験記録システムにおける写真撮影と音声録音の相互補完性

Mutual Complementarity of Photos and Audio in Experience Recording

中蔵 聡哉 Toshiya Nakakura　角 康之 Yasuyuki Sumi　西田 豊明 Toyoaki Nishida

京都大学情報学研究科
Graduate School of Informatics, Kyoto University

Although photography is a valid method of recording experience, you have to interrupt the conversation to take a picture. So when the conversation becomes lively, users don't take it.
When you recorde sound, it is not necessary to operate, but it has less information.
In this paper, we describe the new method to capture the experience.

1. はじめに

我々は今までに, 写真とメモ書きという日常的かつ直感的な手法を融合し, それらの行為を複数ユーザで共有することで, 体験共有コミュニケーションを支援するシステム PhotoChat を開発してきた [伊藤 2007][角 2008]. 写真と書き込みによる記録は, ユーザが注目した瞬間の視点の上にユーザが関心を持った事柄が記録され, 効果的な体験記録手法となっている. しかし, 写真撮影とペンタブレットによる手書き入力は簡単であるものの, ユーザは立ち止まり両手を使う必要があり, 体験が本当に盛り上がっているときには, 記録を残したい瞬間であるにもかかわらず写真や書き込みが行われないことが多い. これは PhotoChat に限らず, 会話と記録のどちらかを行うかを選択しなければならないユーザによる能動的な操作が必要な記録手法全般に共通する問題である.

一方, 能動的な操作の必要ない記録手法の代表である録音による音声データは, それ単体では体験記録として十分とはいえない. 会話の中の指示語や状況が理解できず, 自分の発話であっても何について話していたのか分からないことがある. また, 河村ら [河村 2007] が指摘しているように, 体験データとしての再利用に際して一覧性と検索性の低さが大きな問題となっている.

我々は, この二つの手法を組み合わせ相互補完させることで, よりよい体験記録を行うことができるのではないかと考え, システムを構築して検討を行った. 常時録音することで, ユーザの発話が自動的に記録され, 写真では抜け落ちてしまう会話の盛り上がりを保存することができている. 写真と音声データを結びつけることで, 音声だけでは分からなかったその時のイメージが写真と手書きメモによって補完されることも確認された. 更に, 音声データを記録された時刻が近い写真に自動的に関連付けることで, 写真が音声データのサムネイルとなり, 目的の音声を探す手助けになる可能性も示された.

音声情報を利用することで, 音声データと結びつけるだけではなく, 会話のセンシングを行うことが可能となる [中蔵 2008]. このシステムを用いて, 写真にタグ付けを行った. このことにより「いつ, 誰と話していたときに撮った」写真であるかという, より直感に近い検索を行うことが可能となる.

連絡先: 中蔵 聡哉, 京都大学情報学研究科, 〒 606-8501 京都市左京区吉田本町工学部 10 号館 214 号室, Tel: 075-753-5371, Fax: 075-753-4961, nakakura@ii.ist.i.kyoto-u.ac.jp

以下本論文では, 写真と音声データの相互補完性について検討する. 2 章で目的とする記録の概要について述べた後,3 章で具体的な実装,4 章で利用例を挙げ検討し,5 章でまとめる.

2. 体験共有コミュニケーションの支援

本章では, 節 2.1 で今まで行ってきた写真と手書きメモの融合による体験記録の概要を述べた後, 節 2.2, 節 2.3 で今回実装した音声と写真との組み合わせの概要とマイク入力の性質を用いた会話のセンシング, その必要性について述べる.

2.1 写真と手書きメモの融合

日常的に写真を見ながら他者と会話することは多い. 写真はコミュニケーションを触発することが知られており [Kindberg 2005], 山下らは写真を他者との間の新しいコミュニケーションチャンネルを開くきっかけとして利用している [山下 2001]. このように, 写真は人が何かに注目しているときの視点を切り取ったものであり, それを共有することで, グループ内で互いの興味への「気づき」を得ることができる.

同期の写真共有, すなわち撮影と同時にグループ内で写真を共有することが できれば, 体験の現場で互いの興味を知ることができ, それは非同期の場合とは違った効果をもたらすと期待される. 例えば, 他者の写真を見ることで, 自分が見落としていたことに気づくということがある. 非同期共有の場合はすでにその体験シーンは終わってしまっているため, そこから新たな体験へとつながることはあまりないが, 同期共有により体験の現場で気づくことができれば, その場で反応を返すことができ, 新たな体験創造へつながると期待できる. また, 従来のカメラは後で見返すことが前提となって撮影される場合が多いが, 同期共有が前提になると, その場で他者に見せるために撮影するといったことが起こり得る. 従来のカメラでは撮影されることのなかったようなものが被写体となることもあるだろう.

このように, 写真を共有すると, グループ内のコミュニケーションが触発される. そのコミュニケーション手段として直接会話があるが, 本研究ではさらに共有した写真の上への手書きメモを新たなコミュニケーション手段として用いる.

手書きメモは日常的に行われる記録手段であり, 文字や記号, スケッチなどにより自由な記録が可能である. これを写真と融合することで, 互いに情報強化することができる. すなわち, 写真だけでは撮影者の意図やその時の状況を読み取るのが困難な場合でも, 手書きメモを加えることでそれを明確にすることができる. 逆に手書きメモだけでは十分な表現が困難であった

り時間がかかったりするが, 写真と組み合わせることで簡単化できる. その上, 写真と手書きメモの直感性や手軽さは失われない.

2.2 写真と音声の融合

前節では写真と手書きメモの融合とその同期共有の有効性について述べた. しかし, 写真撮影とペンタブレットによる手書き入力は簡単であるものの, ユーザは立ち止まり両手を使う必要がある. そのため, 実世界での体験が本当に盛り上がっているときには, 記録を残したい瞬間であるにもかかわらず写真や書き込みが行われないことが多い. そうして抜け落ちる記録を補完するための機能が要求されるだろう. 一つの手法として, 動画によって記録し続けることが考えられるが, 常に対象にカメラを向けるのはより手間がかかるし, ウェアラブルカメラを使用するとしても装着の負荷やデータの質・量の面で問題がある.

本研究では, この問題を解決するため, 音声データを用いる. 音声データは, ユーザに操作負荷をかけず, データ量も比較的小さい. 音声データによって, 写真と手書きメモだけでは表現されなかったその場の雰囲気や具体的な会話が音声によって記録されるだろう.

また逆に, 写真が音声データの利用性を高めるのではないかと考えている. 音声データは単独では体験記録として十分とは言えない. この理由として, 指示語の対象等がわからず内容が理解できないこと, 一覧性が低いため再利用し難いことがあげられる. 音声だけでは分からなかったその時のイメージは, 写真と手書きメモによって補完されるだろう. 従来のボイスレコーダでは記憶を頼りに再生箇所を移動させて目的の音声を探すしかなかったが, 写真が音声データのサムネイルとなり, 目的の音声を探す手助けになると期待する.

2.3 音声情報による会話のセンシング

会話を行っている相手とは, 同じ音を聞いているはずである. 例えばある話者の「おはよう」という発言は, 自分のみならず話し相手にも聞こえている.

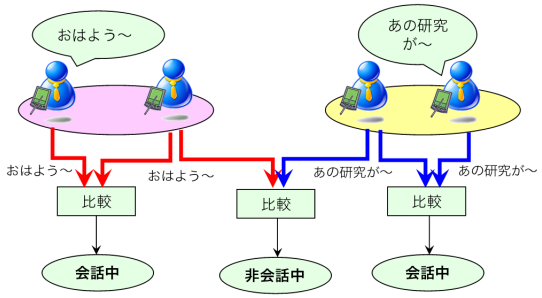

図 1: 音声を用いた会話のセンシング

別の会話を行っている相手は, 同時刻には別の発話者の声を聞いていることになり, マイク入力は全く別の音声的性質を示す. つまり, 入力音声同士を比較することで, 誰と誰が会話を行っているのかを知ることができる.

誰と会話を行っているのかが分かれば, 写真にタグとして付けることで「いつ誰と話していたときに撮った写真であるか」というユーザの感覚に近い検索を行うことができる.

3. 実装

3.1 音声録音

PhotoChat の起動から終了まで、自動的に全ての音を記録する。マイク入力がほとんどない区間を録音しても無駄であるため, マイク入力の音量が小さい無声区間は切り捨て, 図 2 のように無声区間で区切られた有声区間のみを保存する.

無声区間の検出では, マイク入力を 20 ミリ秒ごとのパケットに区切り, パケットの音量がそれまでに入力された音声の音量平均より小さい場合に無音区間と判定する. 200 ミリ秒以上の有音区間があれば録音を開始し, 10 秒以上の連続無音区間があれば録音を停止する.

音量平均は周囲の環境音に強く影響されるため, ユーザの環境に合わせて適応的に変化する閾値となる. したがって, 固定的な閾値では騒音が大きいと全て有声区間となってしまうことがあるが, この方式ならば比較的環境音に影響されずに有声区間と無声区間を判定することができる. その上, 音量平均は次式のような簡単な計算で求めることができる.

$$Avg_n = Avg_{n-1} + \frac{x_n - Avg_{n-1}}{n}$$

ここで,Avg_n が求める音量平均であり,Avg_{n-1} は前回の音量平均である. n はそれまでに入力された音声パケットの数を表し, x_n は n 個目の音声パケットの音量を表す. このように比較的単純な計算で済むため, リアルタイム処理が可能となる.

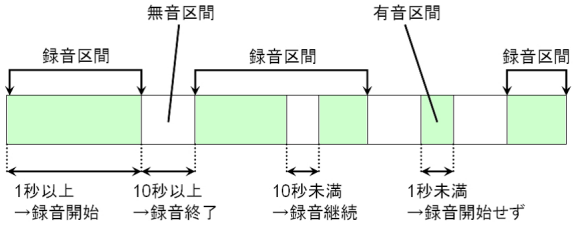

図 2: 無音区間による音声データの切り分け

3.2 写真への関連付け

記録された音声データは, 閲覧・書き込みモードにおいて写真を表示したときに, 自動的に関連付けられてユーザに提示される. 具体的には, 音声データはそれぞれ録音された時刻を保持しており, 写真を表示したときにはその写真の撮影時刻付近 (前 2 分, 後 3 分) に録音された音声データが検索される. 見つかった音声データは, 図 3 のように写真の下に蓄音機アイコンとして時系列に表示されるので, ユーザはそのアイコンをクリックすることで撮影時刻付近の音声を聞くことができる.

このように自動的に写真と関連付けられるので, 撮影時の音声から雰囲気をつかんだり, 聞きたい音声データをサムネイルを頼りに探したりすることが可能となる. 一方で, 音声データが記録されていたとしても, その付近で写真が撮影されていなければ音声を再生する手段がない. したがって, ユーザは後から聞き返したいと思う会話がなされたときには, 少なくとも写真を空撮りしておく必要がある. これは手間がかかることではあるが, 音声データを探すときの目印となる写真をユーザが撮影するよう促す効果があると期待している.

音声データは写真や手書きメモとは違い, 自動的に他のユーザと共有されることはない. これは, 通信データ量とプライバシーを考慮してのことである. しかし, 音声データの中にはグループ内で共有したいものもあるだろう. そこで, 図 3 に示すように共有したい音声データを表す蓄音機アイコンを写真の

図 3: 写真と音声の結びつけ

上にドラッグ＆ドロップして貼り付けることで, その音声データがグループ内で共有されるようになる機能を実装した. これにより, 他のユーザに聞かせたい音声データがあれば, 音声のハイパーリンクを作ることで相手に送信することができ, さらにその音声についての説明を書き込むこともできる. また, 意図的に声を録音してそのハイパーリンクを作成すれば, ボイスメールとして使うこともできる.

3.3 音声情報による会話のセンシング

基本的な考え方は, 入力音声が似ていれば同じと判断するという事である. また, 入力が無音であれば会話が行われていないのは自明であるため比較は行わない. 具体的なアルゴリズムは以下の通りである.

送信側の処理

1. 音声をバッファに蓄積しながら 3 秒待機する.
2. 1. で待った 3 秒の間に有音区間があれば 3. へ. なければ 1. へ戻る
3. 有音区間を最も長く含む連続区間 (1 秒間) の音声をフーリエ変換し, 時刻情報を付加してブロードキャストする. 送信後 1. へ戻る

受信側の処理

1. データ受信するまで待機する
2. 受信パケットのタイムスタンプを取得する.
3. スタンプと同時刻の受信側の音声が無音区間であれば 4. へ. 有音区間であれば 5. へ
4. 無音であれば話していない, すなわち異なると判断できる.1. へ
5. 受信周波数情報と自分の周波数情報を比較する
6. 類似度が閾値以上であれば同じ, 未満であれば異なると判定する.1. へ

比較に用いる周波数帯は人の声の周波数帯と言われる 100Hz～4000Hz で, これを 3901 次元のベクトルととらえてコサイン類似度を求める.

4. 考察

4.1 観察された利用例

今までに, 動物園や博物館での利用, ミーティング, 旅行や日常的な環境での利用と, 様々な目的や環境での利用を行った. その際に記録されたコンテンツでは, 撮影時刻の前後の音声データが写真に自動的に関連付けられ, 振り返りやすいものとなっている.

音声データのハイパーリンク機能を使えば音声データを共有することもできる. しかし, この機能はこれまでの実験では利用されなかった. これは, 音声データは後から振り返るときにその場の雰囲気を知るのに役立つものであり, 現場で活用されることはまず無いからであるとも考えられるが, 音声データの閲覧性の低さや, 過去の音を扱うことにユーザが慣れていないことも大きく影響しているだろう.

なお, 音声のハイパーリンク機能が有効に使えそうな例が観察されている. 例えば図 3 では, コーヒーメーカーの音がうるさかったという事を写真と書き込みで表現しているが, 実際にマイクに音が入っていたためそれを添付している.

写真を見ながらその撮影時刻付近の音声データを再生するのは、紙芝居を見ているようであり, 現在の自分という新しい視点からその時の会話を振り返ることができた. また, 十分に現場の雰囲気を記録できていることが多く, 除夜の鐘を見に行ったときの記録では, 鐘を付く僧達の掛け声や鐘の音が録音されていて, 第三者にとっても面白いコンテンツとなっている.

4.2 音声記録の写真に与える影響

当初述べた, 写真と手書きメモのような能動的な記録手法では会話の盛り上がりが記録されないという問題は, 音声を自動的に録音することによって解決できた. 写真を表示する際に音声データを再生することで, 紙芝居を見るようにその場の雰囲気を理解することができるし, メンバ全員を直接写真に写さなくとも, その場に誰が居たのかも記録されている.

但し, 音声データの質が悪く, 何を話しているのかが分からない事があった. これは, 音質が悪く聞き取りにくいものと, 話し相手の声が録音されていないものの大きく二つに類別される. 音質は高性能なマイクの利用やノイズリダクションの処理を行うことである程度の緩和が可能であるが, 話し相手全員の声を保存するには, 音の減衰といった音響的な問題の他にプライバシーの問題が障害となる. PhotoChat の性質上, 複数人で体験記録を行うという了解が相互にあるため, ユーザ同士の会話であればプライバシーの問題は緩和されると考えているが, PhotoChat を利用していないメンバとの会話をどう扱うか等複雑な問題がある.

4.3 写真の音声記録に与える影響

冒頭で, 音声データの利用性の低さの問題として、大まかなコンテキストが掴めないと何について話しているのか分からないこと, 検索性と一覧性の低さをあげた. これらの問題は, 写真と組み合わせることにより緩和されたと考えている.

まず, 写真と結びつけることで, 撮影された対象からそのとき一緒に居たメンバや場所等を知ることができ, 録音された際の大まかなコンテキストを理解することができ, 音声ログの内容把握自体にも役立てることができた.

また検索性と一覧性の低さは, 写真と結びつけることにより,「除夜の鐘をついた時に話していたこと」等, 写真からその周囲の会話内容を大まかに把握することができた. 逆に, 会話を再利用するために写真を撮影するという行為が観察された. この場合, ユーザが意図的に結びつけているため, 写真のサムネイルを見ることで, 意図した会話を探し出すことができた.

但し, 会話は常に変化するため,「その場に居ない人の噂話」等, その場のオブジェクトとは結び付けにくい会話をどの写真と結びつけるのかという問題がある. この会話をした直後に意図的に写真を取り, その写真の上にメモをしておくというユーザ側の工夫が見られたが, システムが自動的に記録することは現状では難しい.

4.4 写真への会話相手情報のタグ付け

既存のタグ付けでは GPS 情報等が用いられているが, 直前の写真との差異が少なかったり, 地名を思い出さないと使えなかったりと利用しにくい. 今回採用した会話相手の情報をタグとしてつけることで, より直感的な検索を提供することができた.

会話相手がタグ付けされている写真には必然的に会話が結び付けられているため, 会話相手とサムネイルから会話の内容を漠然と知ることができ, 単純に写真をサムネイルにするよりも探索性が高く効果的であると思われる.

5. おわりに

本論文では, 写真と音声の組み合わせによる記録手法の提案を行った. 当初想定していた, 写真と手書きだけでは記録することのできない会話の盛り上がりを保存することが確認でき, 音声録音による写真の体験記録を高められる可能性が示された. 更に, 写真によって, 音声記録に一覧性を持たせたり, 大まかな内容を知ることで音声記録自体の内容把握を助ける等, 音声記録の価値を写真によって高められる可能性もまた明らかになった. 音声情報を用いることで, 会話相手が誰であったかをセンシングすることができ, タグとして付与することができる. このタグ情報は, 写真の検索のみならず, 検索された写真をサムネイルとして利用することで, その会話に参加していたメンバは, 再生しなくとも音声データの大まかな内容を知ることができた.

本論文では, 写真と音声データとの結びつけにおいて, 無音区間で区切り前後の写真と結びつけるという単純な手法をとったが, 熊谷ら [熊谷 2005] の研究では, ポスター発表というドメインでの会話を, ポスターへの指差し行為を用いて会話を構造化し, ポスターと会話の結び付けを行っている. PhotoChat においてもこの手法と同等の工夫を行うことは可能である. 写真と音声を組み合わせた体験記録が, 第三者が追体験を行う場合においてどの程度有効なものであるのかという点とあわせて今後検討していきたい.

謝辞　本研究の一部は情報処理推進機構 (IPA) の未踏ソフトウェア創造事業の補助を受けて行われた.

参考文献

[伊藤 2007] 伊藤 惇, 角 康之, 久保田 秀和, 西田 豊明: 写真と書き込みの実時間共有による学会参加者間のコミュニケーション支援, 人工知能学会全国大会 (第 21 回), 宮崎, 2007 年 6 月.

[角 2008] 角 康之, 伊藤 惇, 西田 豊明：PhotoChat: 写真と書き込みの共有によるコミュニケーション支援システム, 情報処理学会論文誌, Vol.49, No.6, 2008 年 6 月.

[河村 2007] 河村 竜幸, 中西 英之, 石黒 浩: 連続音声録音を用いた会話体験の探索, 人工知能学会全国大会 (第 21 回), 宮崎, 2007 年 6 月.

[中蔵 2008] 中蔵聡哉，角康之，西田豊明: 音環境の類似度に基づいた会話場の認識と利用, インタラクション 2008, A-105 (2008).

[Kindberg 2005] Kindberg, T., Spasojevic, M., Fleck, R. and Sellen, A.: I saw this and thought of you: some social uses of camera phones, Proceedings of CHI '05, ACM, pp. 1545-1548 (2005).

[山下 2001] 山下清美, 野島久雄：思い出コミュニケーションのための電子ミニアルバムの提案, ヒューマンインタフェースシンポジウム 2001 論文集,pp. 261-264 (2001).

[熊谷 2005] 熊谷 賢, 角 康之, 間瀬 健二, 西田 豊明：ポスター発表における発表者と聞き手の間の対話シーンの意味的構造化, 人工知能学会全国大会 (第 19 回), 北九州, 2005 年 6 月.