Vol.2009-UBI-22 No.18
2009/5/15

音声メモの配置・ブラウジングのための胸部装着カメラによる頭部方向推定

山 添 大 丈[†1,*1] 米 澤 朋 子[†1,*1] 寺 澤 洋 子[†2,*1]

本稿では，ユーザの相対的な頭部方向と関連づけて音声メモを記録し，頭部方向に応じて３次元音響として複数の音声メモを提示する音声メモの記録・ブラウジングについて述べるとともに，特に提案システムにおける頭部方向の推定手法について詳しく述べる．胸部装着カメラによる頭部方向推定手法は，頭部へのセンサなどの機器を装着することなく，ユーザの胸部に上向きに装着したカメラの画像のみから，ユーザの頭部方向が推定できる手法である．音声メモの記録・ブラウジングシステムは，音声メモを用いてユーザがホワイトボードの領域の自由な場所にメモを書き込むのと同様な思考空間と，通常の視覚的なメモと同様の一覧性を提供することを目指したものであり，両手法の組み合わせにより，ユーザへの負荷が少なく，より利用しやすいシステムが実現すると考えられる．

Head pose estimation using body-mounted camera for voice-memo allocation/browsing

Hirotake Yamazoe,[†1,*1] Tomoko Yonezawa[†1,*1]
and Hiroko Terasawa[†2,*1]

In this paper, we introduce an intuitive interface of both the recorder and browser of the personal voice memo using user's relative head directions, and especially describe about head pose estimation using a body-mounted camera which is used in the proposed system. The body-mounted camera based head pose estimation can estimate user's head poses without requiring the users to wear cameras or sensors on their head. Our recording/browsing system aims to realize the useful memo space with using the user's voice and head instead of pen-type descriptions such as the white boards or paper memorandums. The combination of our recording/browsing system and the body-mounted camera system overcomes the disadvantages of the conventional devices such as its obtrusive visual impact on other persons and achieves the system that is easy to use.

1. はじめに

複雑な問題について考えたり，複数の問題について整理したりするとき，人はホワイトボードやノート，メモ用紙などに絵や短い記述（メモ）を書き，さらにその記述を一覧したり，並べ換えたりする．近年，携帯電話やＰＤＡ，スマートフォンなどの普及により，このようなメモがどこでも簡単に取れる環境が整ってきている．

しかしながら，このような視覚的なメモ手法は，いずれもメモを記録するための方法・状況が限定的であり，例えば駅でふと何かを思いついてメモしたいと思ったとしても，メモ用紙とペンなど，メモできるものを持っていなければならず，また荷物などを持っていて両手がふさがっていてメモができないといった状況もありうる．

ＩＣレコーダや携帯電話の録音機能を用いた音声メモもよく用いられており，備忘録として入力するだけではなく，作曲家が思い浮かんだフレーズを吹き込むなど多岐にわたる用途があり，音声によるメモの需要は確認されている．一方でこれらの音声メモは，複数個の音声メモを含むときに探索が容易ではない．IC レコーダであれば時間系列に基づいたファイルの羅列であり，すべてを同時に効率的に探索することは音声の性質上不可能である．

これに対し，我々は，ユーザを軸とした相対的空間および相対的方向に着目して，ユーザ発声時の頭部方向情報を用いて仮想的にユーザにとっての相対的空間に音声メモを記録・配置し，それら記録・配置された音声メモがユーザの頭部方向に応じてその存在する方向の

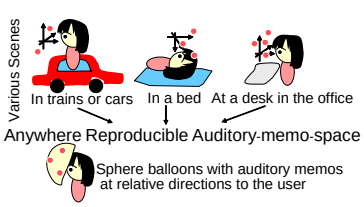

図 1 提案システムのコンセプト

図 2 ユーザに相対的に配置された音声メモ空間

†1 ATR 知能ロボティクス研究所
ATR Intelligent Robotics and Communication Labs.
†2 Stanford University, Department of music, CCRMA
CCRMA, Department of Music, Stanford university

*1 本研究は (独) 情報処理推進機構「未踏 IT 人材発掘・育成事業」の支援により実施したものである．

Vol.2009-UBI-22 No.18
2009/5/15

音源定位で聞こえるように提示する音声メモの記録・ブラウジングシステムを提案した (図 1,2)[1)]．このシステムでは，メンタルスペースのように自由に広がる思考空間することを目指しており，また，人間が持っている聴覚的な分離機能[2)] を利用し，視覚的なメモと同様の一覧性を実現している．

本稿では，特に 1) のシステムにおける頭部方向の推定手法について述べる．1) では，ユーザの頭部方向を推定するために，頭頂部に 3 軸地磁気・加速度センサを装着し，そのセンサデータを用いていた．しかし，日常的に装着して利用することを考えると，ユーザの疲労や周囲への視覚的なインパクトなどの観点から，頭部への装着機器が増えることは望ましくない．

これに対し，我々はユーザの胸部に上向きに装着したカメラを用いて，頭部には何の装置も装着することなく，ユーザの頭部運動を推定する手法を既に提案している[3)]．そこで，本稿では，1) のシステムにおける頭部方向の推定手法について手法 3) を組み合わせて拡張することで，ユーザへの負荷が少ないシステムを目指す．

次節では，関連研究について述べ，３節では，1) 提案した音声メモの記録・ブラウジングシステムについて述べる．４節では，ユーザの胸部に装着したカメラによる頭部方向の推定について説明し，５節で本稿をまとめる．

2. 関 連 研 究

2.1 音声ファイルの配置・ブラウジング

音声ファイルのブラウジングに関する先行研究として Heise らは音源定位を用いたサウンドエフェクトのブラウジングを提案している[4)]．このようなシステムでは，人間の音源定位の聴取機能を積極的に活用している．

一方で，現実世界の対象物や映像コンテンツ上の対象に音声アノテーションをつけるシステムも紹介されている[5),6)] が，これらのシステムはそれぞれの対象が絶対座標に基づいて設定されており，ユーザを中心とした相対的な空間は考慮されていない．

Kanada らは Soundscape[7)] という音声空間の概念に基き，音声コミュニケーションを音響空間内で実現する Voiscape[8)] を提案している．このシステムでも，ユーザが音響空間を意識し，ユーザ自身のメンタルスペースを無意識に反映するように空間との関連付けを行いながら，音声メモを録音するような枠組みは持っていない．

これらのシステムに対し，我々は，ユーザにとっての相対的空間および相対的方向に着目した．ここでの相対的方向とは，胴体の向きに対する頭部方向を意味する．そして，メンタルスペースのように自由に広がる思考空間を実現するため，視覚的にメモ空間を活用しているように，聴覚的にも音声メモを相対的空間に定位しながら録音・配置することや，聴覚的な一覧性を保ちながら録音した音声メモ群のある音響空間をブラウジングすることを目指している点がことなる．

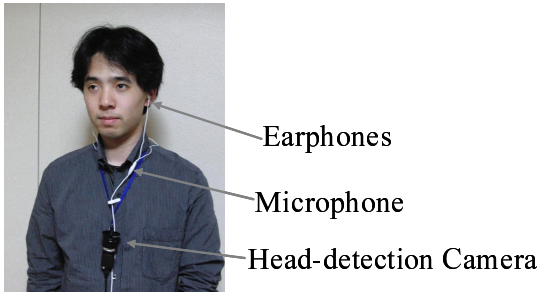

図 3 提案システムの外観

2.2 ウェアラブルカメラによる動作認識

ウェアラブルカメラによるユーザの動作・行動を認識は，これまでに多くの手法が検討されている[9)–11)]。

例えば、Healey らは、'StartleCam' を提案しており[9)]、身体に装着したカメラを用いて、ユーザが驚いた瞬間の画像を取得している。また、ウェアラブルカメラを用いて、ユーザのジェスチャなどの動作の認識を目指した研究も行われており、Starner らによる'Gesture Pendant' では、ペンダントのように首から下げたカメラを用いてユーザの手のジェスチャを認識している[12)]。

ウェアラブルカメラを用いた多くのシステムが検討されているが，頭部運動に着目し，ウェアラブルカメラによってユーザの頭部運動を取得することを目指したシステムはこれまで存在していない．

3. システムの概要

図 3 に提案システムの構成を示す．提案システムは，音声メモの再生・録音用のステレオイヤフォンとマイク，頭部方向推定用カメラ，周辺環境撮影用カメラからなっており，２台のカメラはユーザの胸部に装着されている．

図 4 に処理の流れを示す．処理は音声メモ録音モード（図 4–A）とブラウジングモード

Vol.2009-UBI-22 No.18
2009/5/15

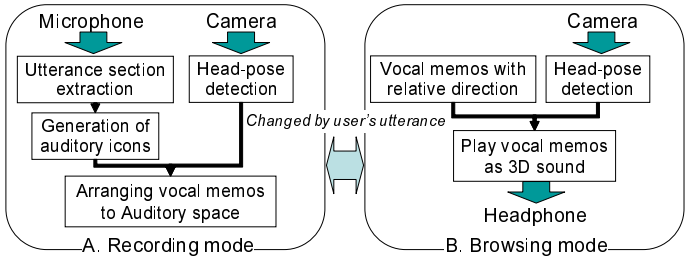

図 4 処理の流れ

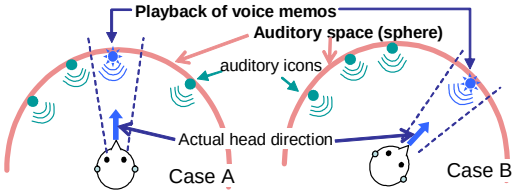

図 5 音声アイコンと音声メモ

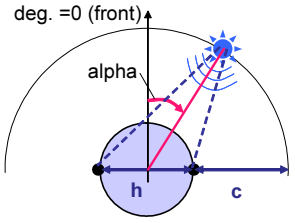

図 6 音源定位の設定

（図 4–B）の２つのモードからなり，ユーザの発声の有無によって切り替えられる．

ユーザの発声区間を検出すると，音声メモ録音モードなり音声メモが記録される．まずマイクで得られた音声データから，音声区間を検出し，音声メモ・音声アイコンを生成する．同時に，胸部に装着されたカメラによりユーザの相対頭部方向を推定し，音声区間における頭部方向をもとに，頭部方向情報付きの音声メモを作成する．

ブラウジングモードでは，現在の頭部方向と音声メモに付加された方向情報をもとに，各音声メモが３次元音声として提示されており，これにより複数の音声メモを同時に把握することが可能となる．ここで，複数の音声メモをブラウジングする際に，音声そのものを複数同時に聞くことは，カクテルパーティー効果を前提としても，音声メモの内容を聞き取ることが難しくなると考えられる．聞き取り対象の音声メモの内容とそれ以外の音声メモの存在をユーザが同時に把握できるようにするため，図 5 に示すように，ユーザの頭部方向の $\pm 5^\circ$ の範囲に存在する音声メモについては，録音した音声メモをそのまま再生し，それ以外の位置に存在する音声メモについては，音声アイコン（抽象化された短い音声）として再生することとし，音の高さが違う鐘の音を用いることとした．

3.1 頭部方向に応じた音声メモの録音

以下では，ユーザの発声区間の検出から音声メモの録音処理について簡単に説明する．発声区間の検出では，閾値以上の音声信号が 0.5[sec] 以上続いた場合に音声発話開始とみなし，その後閾値以下の音声信号が 1.0[sec] 以上続いたときを音声発話停止とみなす．ただし，発話開始判定後 1.0[sec] 以内に音声停止区間が 2.0[sec] 以上続いた場合には，発声区間とみなさないこととする．このような音声区間の検出の自動化が不可能な状況，例えば距離の近い自分以外の発話者が多い等であれば，発話区間判定を手動スイッチなどに切り替えることも考えられるが，ユーザビリティやスムースな入力を想定すると，自動的に音声区間を検出できることが望ましい．

また，一方向を見て音声を入力したつもりでも，ある程度の頭部方向のぶれがあると考えられるため，音声メモ入力開始時や終了時の値ではなく，発声区間中の α, β の平均値をその音声メモの記録方向とする．

3.2 音声メモのブラウジング

ブラウジング時には，現在の頭部方向と音声メモに付加された方向情報をもとに，各音声メモの３次元音声を生成することで，頭部方向に応じた音声メモの提示を実現している．

３次元音声の実装としては，左右１次元的な音声メモ配置のみを考慮しており，左右の音量差および時間差を考慮した単純な処理で３次元の音源定位を実現している．

時間差の計算においては，まず，音源から各耳への距離 dL, dR[cm] を導出する．ここでは，頭部の幅を h, 頭部中心から音源を置く円までの距離を $c+\frac{h}{2}$ として，

$$dL = \sqrt{\{(c+\frac{h}{2})\cos\alpha'' + \frac{h}{2}\}^2 + \{(c+\frac{h}{2})\sin\alpha''\}^2} \quad (1)$$

$$dR = \sqrt{\{(c+\frac{h}{2})\cos\alpha'' - \frac{h}{2})\}^2 + \{(c+\frac{h}{2})\sin\alpha''\}^2} \quad (2)$$

(ただし $\alpha'' = \alpha + 90$)

により計算する（図 6 参照）．そして，時間遅れ $delayL, delayR$[msec] を音速 34 [msec/cm] を用いてそれぞれ $delay = d/34$ により導出する．

音量差については，通常の音源であれば常に同じ音量であっても定位が変わると左右の耳に届く音量 [db] の総合は常に同じとは限らないが，今回の実装では回り込み音声などの計算はしない．よって単純に距離による減衰 atn のみ扱うこととし，$ATN_{init} = 0.8c^2$ として

$$atn = \frac{ATN_{init}}{d^2} \quad (3)$$

により減衰率を決定する．ただし $d_{min} = c$ とする．

Vol.2009-UBI-22 No.18
2009/5/15

4. 胸部装着カメラによる頭部方向推定

これまでに述べたような，ユーザに対する相対的方向に音声メモを配置しブラウジングするシステムでは，頭部方向の検出が必須である．その際，特殊な装置やセンサを装着することなく，音響空間の再現のためのイヤフォンのみを用いることを実現するため，以下では，前節までで述べた音声メモの記録・ブラウジングシステム 1) の拡張として，ユーザの胸部に上向きに装着したカメラによる頭部方向の推定手法について述べる (図 7).

4.1 頭部モデル

まず，ここで用いる座標系について定義する．図 8 に示すように人物座標系 X, Y, Z を定義する．ユーザの視線方向は Z 軸回転には依存しないので，ここでは，X, Y の 2 軸についてのみ考える．α, β はそれぞれ Y 軸，X 軸回りの回転角を表す．人物頭部の回転中心は首の正面 (図 8) に存在すると仮定する．

4.2 頭部領域抽出

赤外照明と赤外カメラを用いることで，人物領域 (頭部・肩部) を抽出する．一般的なモノクロカメラは，可視光領域だけでなく，赤外領域にも感度を持つ．この特性を利用し，赤外照明は人物領域 (カメラ・照明からの距離が近い領域) のみを照らすように調節することで，撮影画像から人物領域を容易に抽出することができる (図 9)．このような赤外照明と (赤外) カメラの組み合わせは手のジェスチャ認識などで用いられている[13)]．

次に，抽出された人物領域から，頭部と肩との境界線 (図 10) を決定する．ここで，頭・肩間の境界線は肩領域によるエッジを直線近似した線として推定できると仮定し，Random Sample Consensus(RANSAC) により頭・肩間の境界線を推定する．

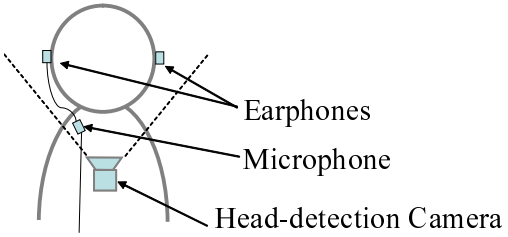

図 7 システム構成

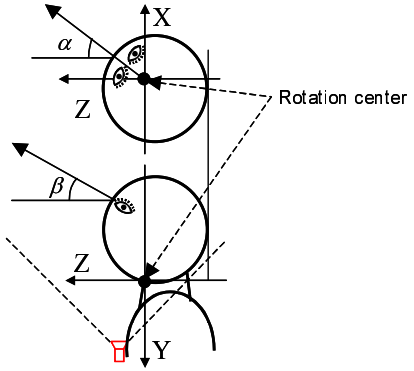

図 8 頭部推定のための頭部モデル

図 9 人物領域抽出

続いて，頭部領域の重心を推定し，頭部重心を通過し頭・肩間の境界線に垂直な線をひく．この垂直な線と頭・肩間の境界線の交点を頭部回転中心とする．また鼻先点は頭部中心からの距離が最大となる頭部領域の点として抽出される（図 10）.

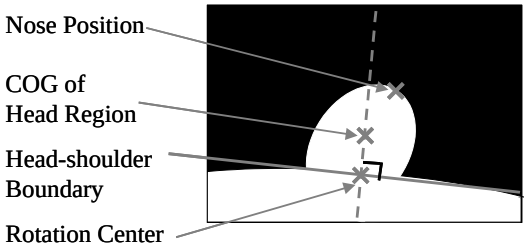

図 10 特徴抽出

以上で述べた処理の結果例を図 11 に示す．ここで，× 印は推定された鼻先点であり，直線は推定された頭部-胴部の境界線である．

4.3 頭部姿勢推定

次に，赤外画像から抽出された鼻先点や頭部回転中心などの画像特徴を用いて，頭部姿勢を推定する手法について述べる．

図 12 に画像特徴間の関係を示す．ここで，赤外画像上の鼻先点と頭部回転中心をそれぞれ $\boldsymbol{P}_n(=[u_n, v_n])$，$\boldsymbol{P}_c(=[u_c, v_c])$ で表す．u 軸は推定された頭部・胴部の境界線と一致し，v 軸は u 軸と垂直で頭部重心を通過する線である．

Vol.2009-UBI-22 No.18
2009/5/15

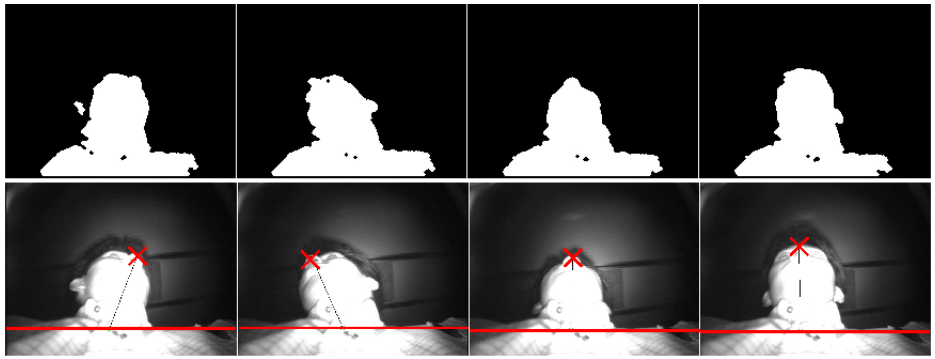

図 11 人物領域抽出 (上: 人物領域抽出結果, 下: 赤外画像と特徴抽出結果)

ここで，頭部回転中心はユーザの首の正面にあると仮定すると，頭部姿勢 α, β は次式で計算される．

$$\alpha = \tan^{-1}\left(\frac{u_n}{v_n}\right),$$
$$\beta = \cos^{-1}\left(\frac{\sqrt{u_n^2+v_n^2}}{k}\right) - \cos^{-1}\left(\frac{\sqrt{(u_n^{(0)})^2+(v_n^{(0)})^2}}{k}\right), \quad (4)$$
$$\left(0 \le \cos^{-1}\left(\frac{\sqrt{u_n^2+v_n^2}}{k}\right) \le \frac{\pi}{2}\right)$$

ただし，k はユーザの頭部の大きさとカメラの内部パラメータにより決定される定数．$\boldsymbol{P}_n^{(0)}(=[u_n^{(0)}, v_n^{(0)}])$ は $\alpha = 0, \beta = 0$ の時の $\boldsymbol{P}_n$ の値．

実際には，ユーザの移動に伴う振動などによりカメラ自体の姿勢が変動するため，カメラと人物間の幾何学的な相対関係も変化する．システムの姿勢変動の影響を吸収するため，頭部姿勢 α, β のオフセット $(\Delta\alpha, \Delta\beta)$ を計算する必要がある．$\Delta\alpha$ は頭部・胴部の境界線の傾き (図 12) であり，$\Delta\beta$ は次式で計算される角度を示す．

$$\Delta\beta = \tan^{-1}\left(\frac{v_c^{(0)} - v_{oc}}{f_{IR}}\right) - \tan^{-1}\left(\frac{v_c - v_{oc}}{f_{IR}}\right), \quad (5)$$
$$\left(0 \le \tan^{-1}\left(\frac{v_c^{(0)} - v_{oc}}{f_{IR}}\right), \tan^{-1}\left(\frac{v_c - v_{oc}}{f_{IR}}\right) \le \frac{\pi}{2}\right),$$

ここで，$\boldsymbol{P}_c^{(0)}(=[u_n^{(0)}, v_n^{(0)}])$ は $\Delta\alpha = 0, \Delta\beta = 0$ の時の $\boldsymbol{P}_c$ の値．f はカメラの焦点距離．よって，補正された頭部方向 $alpha', \beta'$ は次式で計算される．

$$\alpha' = \alpha + \Delta\alpha,$$
$$\beta' = \beta + \Delta\beta. \quad (6)$$

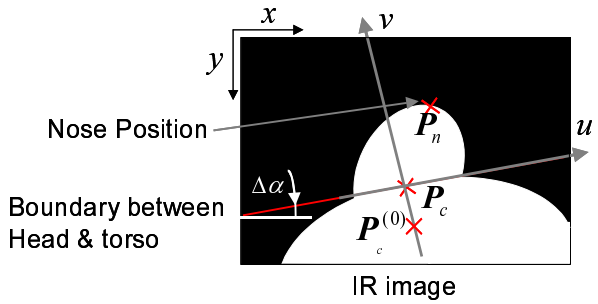

図 12 頭部姿勢推定

以上で述べた処理の結果を図 13 に示す．

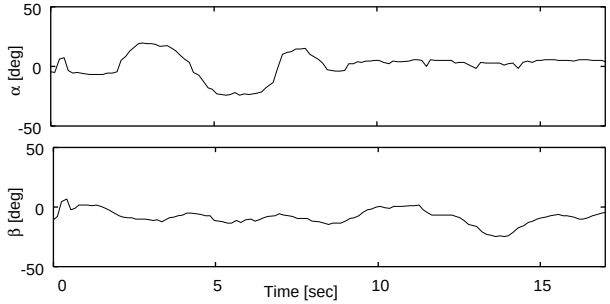

図 13 頭部姿勢推定の例

現在の実装では，胸部に装着したカメラは頭部方向推定だけに用いている．しかし，メンタルスペースの想起・再構築のためには、音声メモを記録した時点でのユーザの環境画像（周辺画像）を取得し，音声メモと対応づけて記録しておくことも重要と考えられる．文献

3) では，胸部に装着した上向きカメラと前向きカメラの２台を用い，ユーザの周辺画像を記録し頭部方向と組み合わせることで，ユーザの注視対象の推定を行っており，上向きカメラと前向きカメラの組み合わせによる周辺画像の取得についても今後，検討をすすめる．

5. おわりに

本稿では，我々の提案するユーザの頭部方向に応じ仮想空間の位置に関連付けた音声メモの録音と，音源定位を用いた複数の音声メモのブラウジング手法について紹介し，本システムにおける頭部推定手法としてユーザの胸部に上向きに装着したカメラを用いた手法について述べた．胸部に装着したカメラを用いて，頭部方向を推定することで，ユーザの頭部にはセンサなどの装置を装着することなく，頭部方向に応じた音声メモの録音や，音源定位を用いた複数の音声メモのブラウジングが実現できるようになる．

今後は、よりロバストな胸部カメラによる人物領域の抽出手法と頭部方向の推定手法を改良するとともにシステムのさらなる小型化などに検討する予定である。また，音声メモの記録・ブラウジングシステムとして，ポータビリティのための追加検討，および，音声入出力関係のハードウェアの検討（音声が雑音に影響されないスロートマイクもしくは NAM マイクの使用や，音像の作りやすいヘッドフォンの開発），より音源定位を感じやすい音声加工上のソフトウェアでの工夫などを検討していきたい．

参 考 文 献

1) Tomoko Yonezawa, Hirotake Yamazoe, and Hiroko Terasawa. Portable recording/browsing system of voice memos allocated to user-relative directions. In *Pervasive 09 Demo*, p. to appear, 2009.
2) A.S. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridgr, MA: MIT Press, 1990.
3) Hirotake Yamazoe, Akira Utsumi, Kenichi Hosaka, and Masahiko Yachida. A body-mounted camera system for head-pose estimation and user-view image synthesis. *Image and Vision Computing*, Vol.25, No.12, pp. 1848–1855, 2007.
4) S.Heise, M.Hlatky, and J.Loviscach. Soundtourch: Quick browsing in large audio collections. In *Proc. the 125th Audio Engineeringg Sciety Convention*, p. Convention Paper 7544, 2008.
5) 垂水浩幸, 森下健, 上林弥彦. Spacetag のアプリケーションとその社会的インパクト. 情報処理学会グループウェア 33–6, pp. 31–36, 1999.
6) Katashi Nagao, Shigeki Ohira, and Mitsuhiro Yoneoka. Annotation-based multimedia summarization and translation. In *International Conference On Computational Linguistics 2002*, Vol.1, pp. 1–7, 2002.
7) R.M. Schafer. *The Soundscape: Our Sonic Environment and the Tuning of the World*. Destiny Books, 1993.
8) Y.Kanada. Multi-context voice communication in a sip/simple-based shred virtual sound room with early reflections. In *Proc. int'l workshopon Network and operating systems support for digital audio and video*, pp. 45–50, 2005.
9) J.Healey and R.W. Picard. Startlecam: A cybernetic wearable camera. In *Proc. of Int'l Symp. on Wearable Computers*, pp. 42–49, 1998.
10) T.Starner, B.Schiele, and A.Pentland. Visual contextual awareness in wearable computers. In *Proc. of Int'l Symp. on Wearable Computers*, pp. 50–57, 1998.
11) B. Clarkson, K. Mase, and A. Pentland. Recognizing user context via wearable sensors. In *Proc. of Int'l Symp. on Wearable Computers*, pp. 69–75, 2000.
12) T. Starner, J. Auxier, D. Ashbrook, and M. Gandy. Gesture pendant: A self-illuminating, wearable, infrared computer vision system for home automation control and medical monitoring. In *Proc. of Int'l Symp. on Wearable Computers*, pp. 87–94, 2000.
13) S.Numazaki, A.Morishita, N.Umeki, M.Ishikawa, and M.Doi. A kinetic and 3D image input device. In *Proc of CHI'98*, pp. 237–238, 1998.