Vol.2015-CVIM-195 No.57
2015/1/23

音響信号との時空間同期を利用した視覚的顕著性の計算モデル

中島 次郎[1,3,a)]　木村 昭悟[2]　杉本 晃宏[3]　柏野 邦夫[2,3]

概要：視覚情報に基づく視覚的顕著性を模倣する計算モデルは，一般物体認識などのフロントエンド処理としての応用が期待され，様々な研究が行われている．しかし，人間の視覚的注意は視覚情報以外のモダリティにも影響を受ける．例えば，人間は大きな音を聞くと，視覚的に顕著でなくてもその音が出ている方向を見る傾向がある．このように，音はしばしば人間の視覚的注意に強い影響を与える．したがって，聴覚情報を視覚情報と併用することで，より高精度な視覚的注意の再現が可能になると考えられる．そこで，本研究では視覚情報と聴覚情報の時間的な同期が視覚的注意に強い影響を与えるという心理物理学的知見に基づき，視覚的注意を模擬する新しい手法を提案する．提案モデルでは，知覚的に顕著である聴覚情報と視覚情報の時間的，空間的な同期を表した時空間同期を逐次的に計算し，同期性を考慮した顕著性マップを算出する．

キーワード：視覚的注意，視覚情報，聴覚情報，Bayesian surprise，視覚的顕著性マップ，視聴覚統合

1. はじめに

視覚情報からの視覚的顕著性の算出，および，それに基づく人間の視覚的注意の推定は，一般物体認識などのフロントエンド処理としての応用が期待され，様々な研究が行われている．例えば，複数の解像度で階層的に画像・動画像特徴を計算し，それをボトムアップ的に統合することによって，周辺との差異が大きく異なる領域を顕著度のマップとして求める計算モデル [1,2] が提案されている．

このような，画像・動画像特徴といった視覚情報を基にした計算モデルの研究は多く行われているが，人間の視覚的注意は視覚情報以外のモダリティにも影響を受ける．例えば，人間は大きな音を聞くと，視覚的に顕著でなくてもその音が出ている方向を見る傾向がある．このように，音はしばしば人間の視覚的注意に強い影響を与える．また，映像には通常，音響信号も付与されており，聴覚情報として音響信号を使用することができる．したがって，聴覚情報を視覚情報と併用することで，より高精度な視覚的注意の再現が可能になると考えられる．

しかし，聴覚情報を利用した視覚的注意に関する研究は十分に行われていない．既存研究の多くは，アプリケーションの精度向上を目的としており，視覚情報と聴覚情報の適合性と視覚的注意の関係には重点を置いていない．視覚的顕著性と聴覚的顕著性を応用した研究の例としては，映像要約 [3] やヒューマンロボットインタラクション [4] が挙げられる．

聴覚的顕著性と視覚的顕著性の統合に関する研究として，音源位置の推定結果を顕著性マップ構成の新しい特徴として組み込む手法 [5] が提案されているが，音響信号を観測してから推定した音源位置を高次レベルで顕著性マップに組み込むにとどまっている．一方，Rolf ら [7] の研究では，視覚情報と聴覚情報の同期性を顕著性マップへ統合する方針がいくつか提案されているが，同期性の検出方法に重点を置いているため，具体的な顕著性マップの構成方法や精度については述べられていない．また，視覚情報と聴覚情報の時間的な同期が視覚的注意に強い影響を与える [8] という心理物理学的知見に基づき，同期性の高い視覚的特徴を選択し，選択した特徴を強調した顕著性マップを構築する手法 [6] が提案されている．この手法は，視覚的注意に影響を与えている特徴を映像を探索してから抽出する 2 パスアルゴリズムであり，逐次的に顕著性マップを計算することができない．また，同期性として重要な画像空間情報を破棄している．

そこで本研究では，音響信号によって強調される画像特徴は，聴覚イベントと同期した時間のみならず，画像にお

1　千葉大学大学院融合科学研究科
2　NTT コミュニケーション科学基礎研究所
3　国立情報学研究所
a)　nakajima13@chiba-u.jp

© 2015 Information Processing Society of Japan

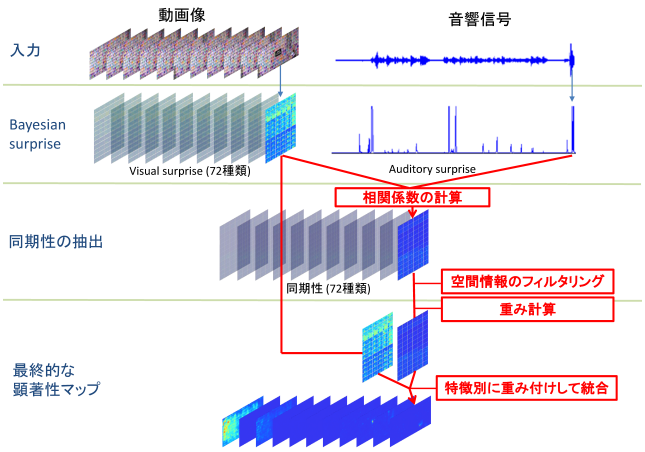


図 1 提案手法の枠組み.

ける空間的な位置にも依存することを踏まえ，時間，空間および特徴別の同期性を考慮した顕著性マップを逐次的に計算する手法を提案する．提案モデルでは，知覚的に顕著である聴覚情報と視覚情報の時間的な同期を画素毎に算出することで，空間的な情報も反映した時空間同期を算出する．その際に，指数平滑法を利用することで時間軸でのフィルタリングを行い，聴覚イベントの知覚と視覚イベント知覚のずれを吸収した顕著性マップを逐次的に計算する．そして，被験者実験から得られる注視点により，提案モデルの性能評価を行う．

2. 提案モデル

本研究では，Bayesian surprise による顕著領域抽出と，指数平滑法による同期性の検出に基づき，同期性を考慮した顕著性マップを構築する．提案モデルの枠組みを図 1 に示す．この図に示すように，提案モデルは 3 段階で構成されている．

2.1 Baysian surprise モデルによる surprise 抽出

第 1 段階ではまず，Itti らの Bayesian surprise モデル [9] により，入力映像から複数の visual surprise map $S_V(t;f,\boldsymbol{x})$ を抽出する．ここで，t は時刻，f は基本となる特徴（輝度，色彩，方向，点滅，運動を基礎とした 72 種）の特徴マップ [10]，$\boldsymbol{x}$ は画素位置を表す．したがって，視覚情報による顕著性として，特徴マップ別の visual surprise map が 72 種得られる．また，Bayesian surprise を音響信号に応用した Schauerte らの手法 [11] により，1 次元の auditory surprise signal $S_A(t)$ を抽出する．

Itti らの Bayesian surprise モデル [9] では，

$$S_{AV}(t;\boldsymbol{x}) = \sum_f \Big(S_V(t;f,\boldsymbol{x})\Big),$$

のように，すべての特徴を加算することで，顕著性マップを構成する．本研究では，聴覚的特徴と時空間的に同期性が高い顕著性を強調するようにして顕著性マップを構成する．

2.2 同期性の検出

第 2 段階では，指数平滑法を使用した同期性の逐次計算を行う．特徴マップ別の visual surprise map $S_V(t;f)$ および，auditory surprise $S_A(t)$ の同期性を，指数平滑法による相関係数の計算方法により計算する．

指数平滑法は時系列分析手法の 1 つで，時系列信号から将来値を予測する手法である．2 つの信号が正規分布に従い生成されているという仮定の基に，音と動画像の特徴の同期性を計算する手法の 1 つとして使用されている [7]．この手法は重み付き平均の計算法の 1 つであり，過去の観測を指数関数的に減少させる重みを使用することで時間方向のフィルタリングを行い，現在の観測付近での平均や分散を逐次的を計算することができる．

指数平滑法では，各時間ステップにおける visual surprise と audiotry surprise の平均値 $\bar{S_V}, \bar{S_A}$ を，

$$\bar{S_V}(t;f,\boldsymbol{x}) = \alpha S_V(t;f,\boldsymbol{x}) + (1-\alpha)\bar{S_V}(t-1;f,\boldsymbol{x}),$$
$$\bar{S_A}(t) = \alpha S_A(t) + (1-\alpha)\bar{S_A}(t-1),$$

で算出する．ここで，$\alpha \in (0,1)$ は指数平滑法による平滑化定数である．同様に，S_V の分散 σ_V，S_A の分散 σ_A，S_V と S_A の共分散 σ_{AV} を算出する．

最後に，分散，共分散 $\sigma_V, \sigma_A, \sigma_{AV}$ から相関係数を，

$$\rho(t;f,\boldsymbol{x}) = \frac{\sigma_{AV}(t;f,\boldsymbol{x})}{\sqrt{\sigma_V(t;f,\boldsymbol{x})\sigma_A(t)}},$$

により算出する．本研究では，相関係数の 2 乗を同期性 $C_{AV} = \rho^2$ とする．

2.3 顕著性マップの構築

同期性を考慮した顕著性マップは，特徴別の visual surprise $S_V(t;f,\boldsymbol{x})$ に対して同期性から得られる重みを反映させることで構成する．中島らの手法 [6] では，入力映像に対する特徴別の同期性を抽出している．そのため，特定のフレーム t，画素 $\boldsymbol{x}$ に対する同期性の情報は破棄されているが，本研究では，各フレーム，画素単位での同期性 $C_{AV}(t;f,\boldsymbol{x})$ を使用して顕著性マップを構成する．

2.3.1 同期性に対する空間情報のフィルタリング

同期性は画素ごとに計算するため，画素間の接続が弱い．そこで，重みを計算する前に，空間情報のフィルタリングを行う．本研究では，特徴マップとガウシアンフィルタによるフィルタリングを行う．

はじめに，特徴マップによる同期性のフィルタリングを行う．特徴マップは中心部に対する周辺部との差異の大きさを画像上に表したものであり，顕著性検出の基礎となる特徴である．この特徴マップの値がある程度大きい位置の同期性のみを使用することで，顕著性検出の元となる特徴の空間的な情報を反映することができる．具体的には，各特徴の特徴マップ $F(t;f)$ において画素平均 μ_F を計算し，

© 2015 Information Processing Society of Japan

平均値を閾値とした 2 値化処理を行い，マスク画像を作成する．そして，同期性にマスク画像を反映させる．

次に，ノイズの除去および，欠損の補間を行うため，マスク画像を反映させた同期性に対し，ガウシアンフィルタを用いたフィルタリングを行う．

これらの処理をまとめると，

$$C'_{AV}(t;f,\boldsymbol{x}) = (F'(t;f,\boldsymbol{x})C_{AV}(t;f,\boldsymbol{x})) * G(\boldsymbol{x};\cdot,\cdot),$$

となる．ここで，F' は特徴マップによるマスク画像であり，

$$F'(t;f,\boldsymbol{x}) = \begin{cases} 1 & \text{if } F(t;f,\boldsymbol{x}) \geq \mu_F(t;f) \\ 0 & \text{othrewise} \end{cases},$$

で算出する．また，G はガウシアンフィルタであり，$*$ は畳み込み演算子を表す．

2.3.2 重みの計算

空間情報のフィルタリングを行った同期性 C'_{AV} に対し，各特徴 f ごとに次の条件を満たすように重みを計算する．(1) 同期性の標準偏差が大きい場合には，平均から離れた値の同期性が全画素に対して多いため，重みに差をつけず，すべて 1 に近づける．(2) 同期性の標準偏差が小さい場合には，平均から離れた値の同期性が局所的に存在すると考えられるため，重みを大きくする．これらの条件を満たす重みは，フィルタリングを行った同期性 $C'_{AV}(t;f,\boldsymbol{x})$ の全画素の平均 $\mu_C(t;f)$ と標準偏差 $\sigma_C(t;f)$ を求め，

$$w_{AV}(t;f,\boldsymbol{x}) = \begin{cases} \exp\left(\beta \frac{C'_{AV}(t;f,\boldsymbol{x}) - \mu_C(t;f)}{\sigma_C(t;f)}\right) & \text{if } C'_{AV}(t;f,\boldsymbol{x}) \geq \mu_C(t;f) \\ 1.0 & \text{othrewise} \end{cases},$$

として算出する．ここで，β は同期性の重み定数であり，値を変化させることでどの程度同期による重みを重視するかを調整できる．$\beta = 0.0$ とすると $w_{AV}(t;f,\boldsymbol{x}) = 1.0$ となり，同期性による影響がなくなるため，通常の visual surprise による顕著性マップとなる．

2.3.3 重みの反映

同期性から計算された重みを使用し，提案モデルの顕著性マップは，

$$S_{AV}(t;\boldsymbol{x}) = \sum_f \Big(w_{AV}(t;f,\boldsymbol{x}) S_V(t;f,\boldsymbol{x}) \Big), \qquad (1)$$

により算出する．

重みを反映させることにより，平均より同期性が大きい場合は顕著性を高め，小さい場合は顕著性はそのままとなる．また，標準偏差が大きい時間帯では重みはほぼ 1.0 となり，通常の visual surprise と近くなる．

3. 注視点による評価実験

式 (1) により出力された顕著性マップを定量的に評価するために，被験者の注視点との一致度を算出した．そして，同期性の高い視覚的特徴を選択するモデル [6] をベースラインとして比較した．The DIEM Project[*1] より配布されている映像（basketball_of_sports）を音響信号とともに 15 人の被験者に提示し，アイトラッカー Tobii TX300 を用いて視線の動きを取得した．この映像のフレームレートは 30fps であり，最初のフレームから 246 フレームを使用して評価実験を行った．観測された視線データで，0.1ms に 2.12 画素以上移動がない視線を注視点として抽出した．精度の評価指標として，Normalized Scanpath Saliency (NSS) [12] を採用した．

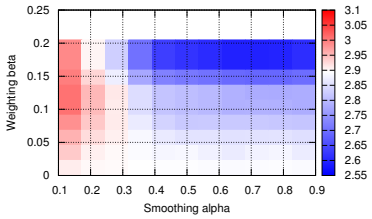


図 2 2 つのパラメータを変化させた際の平均 NSS の値．横軸: 平滑化定数 α，縦軸: 同期性の重み定数 β．

表 1 平均 NSS による精度．

ベースライン	2.896
提案モデル (α, β : 最適値)	3.077

図 2 は 2 つのパラメータ (α, β) による平均 NSS（全フレーム平均の NSS）をヒートマップとして可視化したものである．ここで，赤い領域はベースラインより高い精度を示しているパラメータであり，青の領域はその逆である．ベースラインは，最も高い平均 NSS を示したパラメータでの顕著性マップでの結果を使用している．この図より，$\alpha < 0.4$ の場合において，提案モデルはベースラインと同等以上の精度を有することを確認できる．α は指数平滑法における平滑化定数であり，値が 0.0 に近づくほど過去の影響を重視することになる．したがって，注目フレーム付近を重視して同期性を計算するのではなく，過去の影響を反映させて同期性を計算することで提案モデルの精度は向上したと考えられる．

表 1 に，ベースラインの平均 NSS と，図 2 で最も値が大きい値を示したパラメータ（α, β : 最適値）における平均 NSS を示す．この表から，提案モデルはベースラインと比較して平均 NSS が約 6% 向上していることが確認できる．

図 3 にベースラインと提案モデル（α, β : 最適値）による顕著性マップの NSS の時間変化を示す．ここで，紫と赤の線はそれぞれベースラインと提案モデルの NSS，青の線は auditory surprise を示している．多くのフレームで提案モデルはベースラインより高い精度を示している．

図 4 にフレーム $t = 172$（図 3 においてベースラインと

[*1] http://thediemproject.wordpress.com

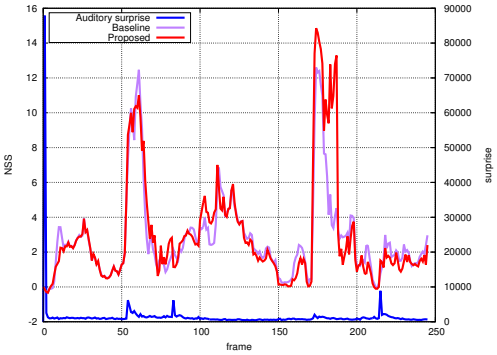
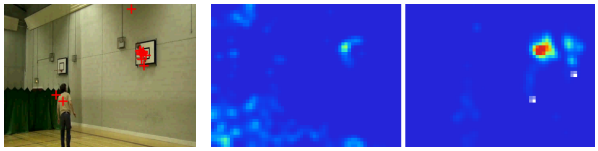
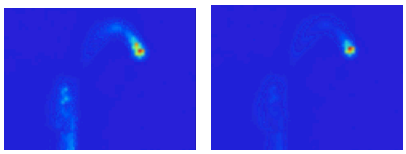


図 3 NSS の時間変化と auditory surprise．横軸: フレーム，縦軸: NSS の値と auditory surprise の値．



(a) 入力映像と注視点．　(b) 重み $w_{AV}(172; f, \boldsymbol{x})$．左: $f = 8$，右: $f = 50$．



(c) ベースライン．　(d) 提案モデル．

図 4 フレーム $t = 172$ における入力映像，重み w_{AV} およびと出力の顕著性マップ．

提案モデルの NSS の差が大きいフレーム）における入力映像，ある特徴における重み w_{AV} およびベースラインと提案モデル（α, β : 最適値）による顕著性マップを示す．ここで，入力映像の赤い十字は注視点を示している．また，特徴の $f = 8$ は色彩特徴（赤青）の 1 つ，$f = 50$ は運動特徴（0° 方向）の 1 つを表している．図 4 によると，提案モデルにより算出された重みは，特徴別に異なる位置や大きさで現れることを確認できる．ここでは，$f = 50$ の重みの画像右上部に大きな値が現れているため，$S_V(172; 50, \boldsymbol{x})$ において右上部の顕著性を強調することになる．顕著性マップにおいて，ベースラインは，画像右上部と左下部に高い顕著性が局在していることを確認できる．また，提案モデルは，ベースラインと同様に顕著性が表れているが，左下部の顕著性はベースラインと比較すると小さい．これは，$f = 50$ ようなの重みを反映させることで，右上部の顕著性が向上し，左下部の顕著性が相対的に低下したためである．したがって，右上部に集中している注視点を，提案モデルはベースラインと比較して高い精度で説明していることを確認できる．

4. おわりに

本研究では，聴覚的な要因によって変化する人間の視覚的注意を推定するための顕著性モデルを提案した．具体的には，Bayesian surprise の枠組みにより視覚的顕著性と聴覚的顕著性を抽出し，それらの時間的，空間的な同期を従来の顕著性マップへ反映させることで，同期性を考慮した顕著性マップを出力した．

また，従来の視覚的顕著性マップと，提案モデルによる視覚的顕著性マップの比較評価を行った．アイトラッカーを用いた被験者実験を行い，その結果を NSS という指標に基づいて定量的な評価を行った．NSS から，時空間的な同期性を利用することで顕著性マップの精度が向上することを確認した．したがって，提案モデルによる視覚的顕著性マップは，より人間の注視を反映させていると言える．

参考文献

[1] Itti, L., Koch, C. and Niebur, E.: A model of saliency-based visual attention for rapid scene analysis, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (1998).

[2] Itti, L. and Baldi, P.: Bayesian surprise attracts human attention., *Vision Research* (2009).

[3] Ma, Y.-F., Hua, X.-S., Lu, L. and Zhang, H.-J.: A generic framework of user attention model and its application in video summarization, *IEEE Transactions on Multimedia* (2005).

[4] Schauerte, B., Kühn, B., Kroschel, K. and Stiefelhagen, R.: Multimodal Saliency-based Attention for Object-based Scene Analysis, *Proceedings of the 24th International Conference on Intelligent Robots and Systems (IROS)* (2011).

[5] Nakajima, J., Sugimoto, A. and Kawamoto, K.: Incorporating audio signals into constructing a visual saliency map, *Proc. Pacific-Rim Symposium on Image and Video Technology (PSIVT)* (2013).

[6] 中島次郎，木村昭悟，杉本晃宏，柏野邦夫：音響信号との同期を手がかりとする視覚的顕著性の計算モデル，第 17 回画像の認識・理解シンポジウム (MIRU2014) (2014).

[7] Rolf, M. and Asada, M.: Visual Attention by Audio-visual Signal-Level Synchrony, *In Proceedings of the 9th ACM/IEEE International Conference on Human-Robot Interaction Workshop on Attention Models in Robotics: Visual Systems for Better HRI* (2014).

[8] Van der Burg, E., Cass, J., Olivers, C. N. L., Theeuwes, J. and Alais, D.: Efficient Visual Search from Synchronized Auditory Signals Requires Transient Audiovisual Events, *PLoS ONE* (2010).

[9] Itti, L. and Baldi, P.: A principled approach to detecting surprising events in video, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2005).

[10] Itti, L., Dhavale, N. and Pighin, F.: Realistic Avatar Eye and Head Animation Using a Neurobiological Model of Visual Attention, *Proc. SPIE 48th Annual International Symposium on Optical Science and Technology* (2003).

[11] Schauerte, B. and Stiefelhagen, R.: "Wow!" Bayesian Surprise for Salient Acoustic Event Detection, *Proceedings of the 38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (2013).

[12] Peters, R. J., Iyer, A., Itti, L. and Koch, C.: Components of bottom-up gaze allocation in natural images, *Vision Research* (2005).

© 2015 Information Processing Society of Japan